Sizing up leadership

Norms and normativity in and around leadership measures

Hesselbo, Emilie

2023

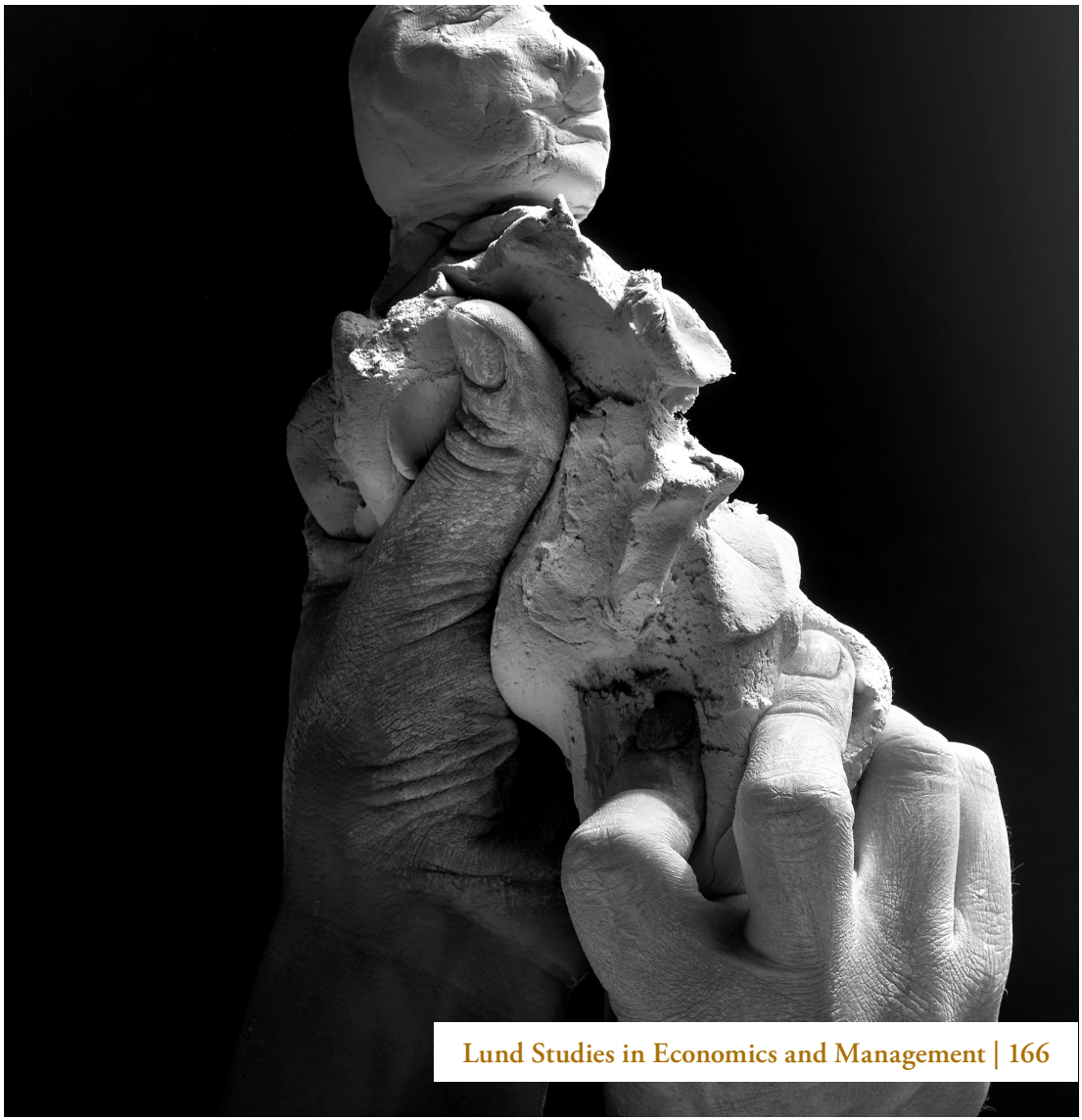
Document Version:
Publisher's PDF, also known as Version of record

LUND UNIVERSITY
SCHOOL OF ECONOMICS AND MANAGEMENT

Sizing up leadership

Norms and normativity in and around leadership measures

EMILIE HESSELBO | DEPARTMENT OF BUSINESS ADMINISTRATION



Lund Studies in Economics and Management | 166

Sizing up leadership

Sizing up leadership

Norms and normativity in and around leadership measures

Emilie Hesselbo

LUND
UNIVERSITY

DOCTORAL DISSERTATION

By due permission of the School of Economics and Management, Lund University, Sweden

To be defended at Ekonomihögskolan April 21st 2023, 14.00

Faculty opponent
Gazi Islam

Organization: LUND UNIVERSITY, School of Economics and Management

Document name: Doctoral dissertation **Date of issue:** April 21st, 2023

Author(s): Emilie Hesselbo **Sponsoring organization:**

Title and subtitle: Sizing up leadership – norms and normativity in and around leadership measures

Abstract:

This thesis explores the role played by norms and social actors in establishing the acceptability and the purported validity of leadership measures. Taking an interpretivist and critical approach, I examine the subjective and normative side of supposedly objective quantitative assessment tools.

Through observations, interviews, and document analysis I uncover the normative agendas and social contexts of four different measurement tools for leadership and personality assessment. I deploy two concepts – *normalising potentials* and *mediating strategies* – to argue that we should understand the performative effects of quantitative assessment tools in relation to test practitioners' and test takers' interaction with them.

Leadership measures and personality tests have normalising potentials, the actualisation of which depends on their broader context, as well as the norms test practitioners mobilise and the mediating strategies they employ. The interaction between hard statistical norms and soft mediating norms is critical for understanding how measurement tools come to have normalising effects on test takers.

These insights extend and add to the existing critical literature on quantitative measures by refocusing attention away from the tools themselves, their powers and effects, to the work and influence of social actors in organisations who develop, frame, sell, present, receive, and interpret the instruments. Future studies on quantitative assessment tools should thus consider the social context surrounding such measures and the mediating work on which these measures' performative potential relies.

Key words: Leadership measures; personality tests; quantitative assessment tools; commensuration; quantification; norms; normalisation; performativity; leadership development

Language: English **ISSN and key title**:

ISBN:
978-91-8039-551-9 (print)
978-91-8039-550-2 (digital) **Number of pages**: 210

Signature  Date 2023-03-09

Sizing up leadership

Norms and normativity in and around leadership measures

Emilie Hesselbo

LUND
UNIVERSITY

Coverphoto by: Sarah Hesselbo

School of Economics and Management | Organization | Lund University

ISBN 978-91-8039-551-9 (print)
ISBN 978-91-8039-550-2 (digital)

Printed in Sweden by Media-Tryck, Lund University
Lund 2023

Media-Tryck is a Nordic Swan Ecolabel certified provider of printed material. Read more about our environmental work at www.mediatryck.lu.se

MADE IN SWEDEN

Table of Contents

Acknowledgements

This thesis is written by me, but made possible and influenced by many different people.

First of all, thank you Sara Louise Muhr for passing on the call for a PhD student. It led me to Lund University and, most importantly, to my very first meeting with what would turn out to be two of the most important and valued people in my life: my supervisors, Sverre Spoelstra and Nick Butler.

It is not easy for me to convey the depth of my gratitude to you, at least not without getting a little emotional. So perhaps this is the perfect form of expression.

I feel so privileged to have had the two of you by my side during my years as a PhD student. And by my side you have been. You have always made yourselves available, reached out a hand when I needed one, and reassured me when feelings of (self-)doubt almost knocked me over. Thanks to you and your thoughtful advice, I have been able to navigate the uncertainties, challenges, and opportunities that academia offers, in the best way possible.

Thank you, Sverre and Nick, for your graceful guidance, encouragement, inspiring commitment, and, most importantly, for your support. I have especially cherished your unique ability to listen, to whatever and whenever. I treasure all that you have introduced me to and taught me, from intellectual perspectives to dangling modifiers.

Then, I would like to extend a big thank you to my opponents and colleagues who have shared with me many valuable insights and reflections. Peter Svensson, Nicole Ferry, Monika Müller, Sanne Frandsen, and Jens Rennstam, you have all contributed with feedback that has sharpened my gaze immensely.

I also owe a thank you to my fellow PhD students, I have had the pleasure of getting to know over the years. In particular, Anna Sophie Biering-Sørensen, Johan Jönsson, Jonas Cedergren, Nina Singh, and Oskar Christensson. Being able to share my frustrations and joys with you has made this endeavour a lot less lonesome. A special thanks to Anna Sophie Biering-Sørensen, my PhD mentor,

office mate, but most importantly, friend. I have valued your company so very much. And to Johan Jönsson, for repeatedly engaging with my material and ideas. Thank you for your comments, insights, and friendship.

Completing this journey has only been possible due to the amount of support I have received from friends and family. I would especially like to thank Sarah Hesselbo, who is not only my much-loved sister, but also the talented photographer who created the front cover of this book. Thank you for helping me realise the vision I had.

Most of all, I am deeply indebted to you, Alexander, and so incredibly thankful to have shared this entire wild ride with you. Your unwavering belief in me, care, comfort, and open arms have made all the difference. Thanks for listening, and then listening some more. For celebrating my accomplishments and cheering me on during setbacks.

And lastly, thank you Asger for keeping the fire burning within me. And around me.

Introduction

How much are numbers worth? When it comes to peanut butter, if you want to know the precise numerical characteristics of every single molecule, protein, and trace element in your sandwich filler, then you can expect to pay $1,069 (for three jars, mind you). As James Vincent, author of Beyond Measure: The Hidden History of Measurement (2022), explains, this costly price of what is known as Standard Reference Peanut Butter is due to the effort involved in mapping and measuring every constituent element of the contents (Vincent, 2022).

When something (or someone) has been quantified in some way, we are inclined to think that we now know something essential about that thing, about how it works, its qualities, and its worth. We value and rely on numbers to such an extent that the absence of numbers in evaluative descriptions, decision making, and monitoring processes are the exception rather than the rule. At the same time, we have a tendency to simply accept the numbers we are given rather than questioning their origin, accuracy, or relevance. For example, as Vincent (2022) shows us, the recommendation that we should all walk 10,000 steps a day is based on the name of a Japanese digital pedometer from 1965. This instrument for counting steps was called the '10,000 steps meter' (*manpo-kei*), with the number in question apparently being chosen because the character for '10,000' is '万', which visually resembles someone or something taking a step forward. Despite the more or less accidental establishment of 10,000 steps as a standard for daily exercise, this number has gained considerable power and influence as a marker of healthy living.

Justifications grounded in numbers generally invite less questioning than do non-numerical arguments. This elevation of quantification is the result of processes that span centuries. Quantitative measures and assessment tools are so prevalent today in both the organisational and the private spheres that Mau (2019) quite reasonably describes our society as a 'metric society'. The scope of social measurements has expanded so much over the past thirty years that it now includes everything from human behaviour and well-being to blood pressure, menstrual cycles, sleep, performance, productivity, personality, and leadership.

Through quantification, we seek to track, evaluate, and optimise any phenomenon that we can describe in numerical terms.

The attempt to measure and compare more and more aspects of our lives has reached such a level that some speak of an 'evaluation cult' (Mau, 2019, p.82), a cult that venerates measurements, evaluations, and continuous comparisons. Societies are more and more driven by data, fixated on figures, and inclined to favour numbers above all else (Muller, 2019). With self-tracking devices, personality tests, and endless questionnaires and surveys, (self-)monitoring activities have increased, potentially changing 'how we understand our selves, what we attend to, and how we organize our lives' (Mennicken & Espeland, 2019, p.224).

Given this societal commitment to quantitative analysis, it is unsurprising that the pursuit of evidence-based actions, transparency, and accountability, and the use of quantitative measurement tools have become common practice in many areas of organisational life. For instance, 80% of Fortune 500 companies and 75% of The Times Top 100 companies use psychological tools for hiring and promotions that target abilities, knowledge, attitudes, or personality (Psychometric Success, 2022, n.p.). The use of quantitative assessment tools to assist in making decisions about people (e.g. which candidate will get the job in a recruitment process) is widely accepted in Personnel departments (Butcher, 2010). The tools are justified by their promise to add value, bring financial benefits, empower, inform, and increase self-awareness and performance (e.g. Datta, 2015; Melamed & Jackson, 1995). These measures are sold commercially by consultant agencies as a necessity for enhancing organisational performance.

Empirically, this thesis focuses on one area in which such instruments are commonly used, namely leadership development. This is a field in which consultants, or those who develop and conduct the tests (henceforth 'test practitioners'), try to capture, map, and improve leadership potential or effectiveness. Reflecting a leader-centric orientation, according to which self-awareness and self-reflexivity is key to leadership effectiveness, assessment tools are used to target the individual's 'inner self', promising to provide such benefits as 'insight' and a 'deeper understanding of character' (Sigma Assessment Systems, 2023, n.p.). Advocates for the tools claim that the instruments have 'the ability to unlock an organization's long-lasting potential' (Unique HR, 2016, n.p.). This view, together with the notion that 'unless something can be measured, it cannot be improved' (Kelly 2007 cited in Moore & Robinson, 2015, p.7), has led

leadership development programmes to increasingly make use of quantitative assessment tools such as personality tests and 360° assessments.

Tests and measurements have thus become a popular, naturalised, and taken for granted activity in organisations and leadership development programmes. Their format is alluring and appeals to our preconceptions about the accuracy and trustworthiness of quantitative knowledge, especially when compared to 'merely' qualitative results. As Porter (1995) states, the claim made on behalf of numbers is that the 'desires and biases of individuals are screened out' (p.74), purportedly producing credible, objective knowledge. The status attributed to quantification enables leadership measures to speak to organisations' need for hard data by which to ground their decision making. Consequently, quantitative measurements and assessment tools are perceived to be an objectively superior solution for most organisational challenges. Test practitioners claim in quite causal terms, how the instruments create order and predictability and ensure objective and fair assessments. Strengthening the conviction that quantitative measures are preferable when making assessments in organisations, test developers and users tend to place them in direct opposition to 'subjective' processes relying on 'gut instinct'.

The attempt to measure leadership numerically follows a belief that leadership has a quantifiable form. Tests and measures represent the assumption that there are clear and measurable dimensions that together constitute a person's abilities as a leader. A Google search for "leadership tests" generates thousands of hits including endless quizzes and questionnaires encouraging one to: 'Take the leadership test and see if your personality traits lend themselves to a successful leadership role' (Psychometric Tests, 2013, n.p.), with pop-ups urging you to 'Invest in yourself'. Other test sites promise to tell you 'whether you possess the personality traits and skills that characterize good leaders' (MindTools, 2022, n.p.). On such sites, the predominant assumptions are that personality traits are measurable and have predictive strength, and that knowledge exists about what constitutes successful leadership.

Measurements and related evaluation systems represent certain expectations and prescriptive standards that seek to guide our actions, decision making, performance, and potential self-perceptions in the quest to achieve 'numerical excellence' (Mau, 2019, p.176). Tools which are supposed to measure leadership are therefore not neutral. Besides being based on a normative sample against which test takers' scores are considered and interpreted, tests are constructed on assumptions about which traits and behaviours give rise to effective leadership.

For instance, on mindtools.com, we find the statement that: 'Successful leaders tend to have certain traits. Two key areas of personal growth and development are fundamental to leadership success: self-confidence and a positive attitude' (2022, n.p.).

Indeed, leadership measures are not value-free. The instruments are informed by and built on norms and there is therefore a tendency for them to guide the test taker towards the right personality and a 'successful' leadership style. Most significantly, tests are used with the aim of improving something. Therefore, what the results imply about suggested behaviour change reveals strong prescriptive standards. In other words, leadership measures rely on (re)producing and sustaining certain beliefs about good/bad leadership. Leadership constructs such as *authentic leadership, transformational leadership, servant leadership,* or *spiritual leadership* all represent certain convictions about good/bad leadership, and so the measures developed to improve these leadership styles do so as well.

Since leadership is a value-laden concept, often assigned vital importance by both scholars and organisations, the stakes are high when one attempts to measure it. One's leadership abilities, or others' perceptions of them, can lay the foundations for possible promotions, access to talent programmes, improved salary levels, and many other conventionally desirable outcomes. Further, the development of good leaders is seen as an imperative for organisational success (e.g. Bass & Avolio, 1993; Dalakoura, 2010; Leskiw & Singh, 2007) on the grounds that 'one of the best ways to grow … organizations is to develop their leaders' (Mehrabani & Mohamad, 2015, p.821). Successful leadership is considered to be critical for organisational performance, which is why organisations strive to cultivate and propagate what is thought to be the optimal leadership style or approach. Since leadership has gained this vital status and importance, leadership development is often considered a high-profile activity and a key element in competitive strategies, meaning that the financial investment is significant (Becker and Huselid 1998 cited in Mabey, 2013, p.359). Indeed, inherent in leadership development are behavioural, reputational, and financial goals, which is why measuring leadership in particular heightens the stakes. However, what the behavioural prescriptions in such psychological assessment measures cover, and how they have come to dominate and influence belief systems, norm constructions, and even perhaps self-perceptions, are tacit and taken for granted. The appropriate leadership style and associated norms simply become the established discourse, risking little or no reflexive contemplation.

Problematization

What is remarkably absent from the debate on quantitative measurements are the social dimensions of the use of such tools (Espeland & Stevens 2008): the individual interpretations and the assumptions, beliefs, and norms that permeate test practice. Moreover, there is a lack of transparency around the tools and their uses that Wilson, Lee, Ford and Harding (2020) describe as no less than 'disquieting' (p.9), since the knowledge produced by quantitative assessment tools is both 'personal and sensitive' (p.9). In the pursuit of furthering our knowledge of the 'ethics of metrics,' Islam and Greenwood (2022) encourage researchers to 'pay more attention to the messy world of practice' (p.4), since it is in practice that the boundaries between quantitative tools' development, use, purpose, and consequences are blurred. In another paper, Islam (2022) argues more specifically for ethical considerations concerning the different processes involved in quantification. These processes entail choices about what to quantify (and as a consequence, what not to quantify), how to commensurate this, and what then happens when numbers are mobilised, deployed, and 'become the property or capital of specific actors' (p.201). In my study, I respond to these concerns by bringing to the centre the social aspect of numbers. I emphasise the choices and influence of social actors, i.e. test practitioners and test takers, at different stages of a measurement process, and highlight the context in which four different measurement tools operate.

More specifically, this thesis is concerned with the norms related to measures and with the social actors who frame, mediate, and interpret the tools. In examining these issues, my study brings to light how social actors and norms work to form the experience of test takers, including their expectations about being measured and their emotional responses to this measurement. I will argue that test practitioners employ strategies that actualise the norms that are built into given measures. By framing, tweaking, foregrounding, or repressing, that is, by mediating numbers in a variety of ways, practitioners establish the necessary conditions for the measures to have an effect on the test taker.

My goal in this study is to draw back the curtain to reveal the veiled side of leadership measures. I inquire into the rationale behind the use of such tests, and the capabilities we ascribe to these measures. Through a study of social actors' strategies, reactions, and interpretations, I examine the work that leadership measures require in order for them to have performative effects. I question the power and performative capabilities that are routinely associated with these

measures, developing an interpretation of the practice that departs from those propagated by test advocates, and extends those of tests' critics. According to this interpretation, testing is an intersubjective and normative process that requires social actors' mediating activities in order for the process to result in effects on test takers' attitudes and behaviours.

In challenging the widespread view of quantitative assessment tools as rather unproblematic and value-neutral, I draw attention to the paradoxes and ambiguities inherent both in the instruments themselves and in the activities and language that support them. Overall, this thesis contributes both to the development of a more reflexive leadership practice and to the critical discourse on quantitative assessment tools, by questioning contemporary trends and revealing their origins, development, and ephemeral nature.

The study I have conducted is primarily a study of the increased use of quantitative assessments and all the activities and efforts that affect them, support them, strengthen and work against them. Leadership and norms serve as the guiding themes of this thesis and have provided a continuous frame for my study. Leadership measures are the object of study, meaning that I give attention to leadership discourse, adding nuances to how we understand, value, and evaluate leadership. With that said, leadership takes both central and more peripheral roles throughout the thesis. The fact that I am studying leadership measures is of importance when it comes to the normativity of the instruments and their inherent goals. When measuring leadership, particular objectives become relevant and certain benefits are at stake (e.g. promotions and access to talent programmes). However, some of the mechanisms and activities surrounding the leadership measurement process are not necessarily unique for leadership measures. My study shows the significance of test practitioners' roles, measurement framing, norms, and individual experience; aspects that might also be significant and distinct in other social measurement processes.

Outline of chapters

The thesis is divided into six chapters.

In the first section of the literature review: 'Quantification, objectivity, and normativity', I present a historical account of numeracy, quantification, and objectivity, while also exploring the different meanings of the term 'norm(al)'. I

show here how the contemporary use of measures relies on historical developments and changes in belief systems, with the emphasis on *belief*. What is important here is the association of quantitative assessment tools with objectivity and rationality. This association rests on faith in the methodology and logics of the belief system in question, as well as in its foundational axioms, such as that personality, leadership, or behaviour are phenomena that lend themselves unproblematically to being quantified. Since quantification, objectivity, and leadership are all based on norms and promote certain normative standards, I explore the concepts of 'norms' and the 'normal' at the end of the section.

In the second section called 'Social measures and their normative implications', I further develop the normativity associated with social measures. I direct my gaze more specifically towards psychometric instruments and leadership measures, including the development of personality tests and leadership discourse. The question informing this review of literature is how leadership has been conceptualised, studied, and measured, particularly throughout the past century. In this review, I highlight how the notion of leadership carries different meanings, norms, and values, all of which affect how people attempt to measure it. I include in this section an overview of the leadership/personality test industry, contextualising and positioning the four measures I have studied.

In the method chapter, I present my methodological choices and reflections. I lay out my philosophical considerations, the choices I have made regarding data collection, my analytical strategy, and reflections on particular challenges and opportunities I have encountered in the field. More specifically, I account here for the methods I have employed (observations, document analysis, and interviews) and discuss their contributions.

From here I move on to my empirical analysis. This chapter of the thesis consists of three sections. In the section called 'Normalising potentials' I shed light on the four tools' inherent potentials to normalise and regulate test takers' attitudes and behaviours. I highlight the particular value-laden language permeating texts related to tests and the visual expressions of the measures' results. The next section: 'Mediating strategies', explores how test practitioners introduce, talk about, and frame the measures in ways that are essential for the tools' normalising potentials to be actualised. In the third empirical section: 'Responses to normalising potentials and mediating strategies', the perspective shifts to that of test takers – including myself – outlining and reflecting on different test experiences and responses, responses that promote the need for mediating strategies, thus tying together the activities surrounding test use.

In the discussion chapter, I re-engage with the literature presented in the literature review through a discussion of my empirical findings and analysis. Through the concepts of *normalising potentials* and *mediating strategies* I unpack how my empirical findings contribute new insights into measurement practice and the performativity of numbers. Drawing on philosopher John Langshaw Austin (1962), I argue that measures, like speech acts, contain performative intents or potentials to normalise, the actualisation of which relies on social actors creating the appropriate circumstances, for example by employing mediating strategies.

The thesis ends with a conclusion in which I reflect on my study's implications in terms of how we approach and think about quantitative measures used in leadership development programmes, and how my findings invite further research that will broaden and nuance our knowledge and use of quantitative assessment tools. By employing the proposed concepts (*normalising potentials* and *mediating strategies*), we can refocus both scholarly and practical attention from quantitative tools themselves to the efforts of their supporters to have these tools accepted as legitimate, valuable, objective, and sources of accurate results.

LITERATURE REVIEW

In this chapter, I trace the historical developments of quantification, its relationship with objectivity, and its infiltration into the discourses on psychology and leadership, with norms as an overall frame. These developments are important for the subsequent analysis of leadership measures in their interaction with test practitioners and test takers.

The chapter consists of two sections. In the first section, I explore how quantification has developed since the seventeenth century and become a well-established method of collecting and handling information. The second section deals with specific examples of quantitative measures such as psychological tests, leadership measures, and their normative implications.

The initial historical account shows us how the use of numbers and quantitative measures have expanded, and most importantly allowed us to look at the world both statistically and strategically: Transforming information into numerical units invites generalisations, comparisons, identifications of patterns, increases, decreases, and deviations. I then examine quantitative measures' relationship with objectivity and rationality, since this relationship is partly what lays the foundation for quantitative assessment tools' perceived legitimacy and status.

A historical account of quantification provides an important context for my study of contemporary quantitative assessment tools. Knowing more about the historical roots of quantification contributes to understanding why the fascination with tests and measurements prevails today and what developments have informed quantification's status and attributed qualities. In particular, my study shows that test practitioners' efforts to mediate measurement activities and experiences utilise the historically forged link between quantification and scientific objectivity. Understanding the development of this link means that we can better grasp and challenge what tests and their developers leverage in their legitimisation efforts. My study thus adds nuances to the role and status of quantification today and its relationship with objectivity, subjectivity, and normativity.

Since leadership measures and personality tests consist of norms, an analysis of such measures unavoidably must consider the different levels of and roles played by norms. More specifically, leadership measures consist of statistical norms and norms about what types of leadership behaviour or levels and combinations of characteristics are desirable. At the end of the first section, I therefore provide an account of the different meanings of 'norm' and 'normal', and how they contain both statistical and behavioural components.

In the second section of this chapter, I turn my attention to more concrete examples of quantitative measures: psychological tests and leadership measures, and their normative implications. The value-laden dimension of quantitative measures becomes particularly relevant when we start measuring our innermost selves – exactly what is targeted in psychological tests and leadership measures. I therefore look into the histories of psychological tests and how quantitative measures have normative implications that guide our behaviour, self-understandings and what we perceive as normal and desirable.

From that, I move on to explore how these normative tests and measures are used in the organisational sphere, and more specifically, in leadership development. This is followed by an outline of the test industry, including the different types of psychometric tests on the market, illustrating the scope and variety of the phenomenon and offering a context for the measures of interest in my study. The second section ends with a review of previous studies on psychometric test use in organisations, both quantitative and qualitative studies – positioning my own research by framing what my study challenges and what theoretical field it contributes to.

Quantification, objectivity, and normativity

A historical account of quantification

Numeracy, the ability to reason, identify, understand, and apply numbers to everyday situations, is a basic mathematical skill needed to quantify and commensurate; the equivalent to literacy. A historical understanding of numeracy contributes to uncovering the current political, societal, and psychological attraction of *quantification*, and more specifically *commensuration*: the act of

turning qualities into quantities e.g. observations or characteristics (Espeland & Stevens, 1998).

Counting, measuring, and testing have existed for as long as the Arabic digit system[1], implying that the inclination to transform phenomena into quantifiable units has a long history. Nevertheless, quantitative measurements have not always been as prevalent in society as today. According to the Old Testament, David brought a plague on Israel for 'numbering' the people (Cohen, 1999, p.35). The religious belief of the 'sin of David' made the English resistant to and sceptical of censuses. Thus, the plan to take a census in England in 1753 was received with protests: the Whig party claimed that it would ruin the 'last freedoms of the English people' (Desrosières, 1998, p.24).

However, what was deemed appropriate to count or quantify varied between countries. Censuses were conducted in 1672 in Holland and in 1749 in Sweden (Desrosières, 1998, pp.24–25). The rejection of censuses in England was perhaps related to a general scepticism. The few educational institutions existing in the seventeenth century ignored arithmetic, considering the skill to be of no value, 'a vulgar study', and not contributing to the primary subject of theology (Cohen, 1999, p.118). Basic arithmetic training was considered too difficult for small children to learn. Nevertheless, in the eighteenth century in North America, older boys requested to learn it and evening classes became available. Arithmetic remained a niche skill, not considered relevant or possible for the many to master. Numeracy was a prestigious skill; its exclusivity was sustained by the incomprehensible language in books written on the subject. In this way, people were discouraged from learning the skill and the spread of numeracy was constrained (Cohen, 1999).

This resistance and scepticism concerning numeracy changed over time. Throughout the seventeenth century in particular, changing societal needs, living conditions, colonisation, the increasing size of countries, commercial capitalism, and overseas trading caused numeracy and arithmetic to gain attention and value, albeit expressed differently in different countries (Cohen, 1999; Desrosières, 1998; Lazarsfeld, 1961).

In England, the concern was with 'quantified objectification' (Desrosières, 1998, p.30). This was the act of quantifying phenomena such as mortality into calculable units or objects. Techniques of recording and calculating were

[1] The ten digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

developed and the term 'political arithmetic' was coined (Desrosières, 1998). Political arithmeticians produced written records of, for example, baptisms, marriages, and burials, which provided the government with information about the lives of individuals, contributing to the administration of the state. For instance, information about mortality rates served as basis for establishing life insurance premiums, and estimates of population sizes in different regions were key when collecting taxes or enlisting soldiers (Desrosières, 1998).

In Germany during the seventeenth century, the ground work for descriptive statistics took place (Desrosières, 1998). Less concerned with quantified objectification, German statistics suggested comparing descriptions of communities (states, regions, towns, or professions). This eventually led to the use of tables and cross-references. In such tables, countries would appear in rows and the different elements of their description in columns. At this point, elements were literary and not numerical, but the form still allowed simultaneous assessment of communities or countries. This was a complete contrast to written or oral material (Desrosières, 1998).

The tabular way of presenting information laid the foundations for quantitative statistics. The form encouraged numerical instead of literary comparison. Placing empirical observations in rows and columns enables identifications of patterns, tendencies, growth, progress, and decline, something that simply listing items, episodes, or names of people cannot (Gregory, 2013). As Gregory (2013) exemplifies, with early demographer and statistician John Graunt's (1620-1674) recordings of deaths during 'plague-time' in London in the seventeenth century, a shift in both focus and knowledge happens when moving from lists to tables. Registering casualties in lists displays individual cases, including names, addresses, and causes of death. When transformed to tabular form, the data is stripped from detail until it appears in 'naked' form, suitable for a numerically organised table (Gregory, 2013, p.313).

This transformation has implications for how one reads lists and tables. When reading a list, one is usually looking for something particular, for example a person, or an overview of a phenomenon such as the multiple causes of death. Reading a table, where details are removed and replaced with a number, one might observe general increases or decreases, pay attention to sums and totals, or quantitative deviations (Gregory, 2013). In tables, information is transformed into a different type of (what appears as) structured knowledge, allowing for the world to be looked at statistically *and* strategically.

Although different in approach and technique, what the English termed 'political arithmetic' and the Germans called 'descriptive statistics' shared was their use for explaining complex phenomena in simpler ways. Whether the purpose was comparison, governing or prediction, both countries worked to reduce and present information in ways different than the common written or oral form, for example through tables and lists.

The two-dimensional form of tables requires common criteria on which comparisons can be based, decided by the 'classifying authority' (Gregory, 2013, p.311). This mechanism has led to a range of objections. Critics argue for example that the tabular medium reduces complex events or phenomena so that they lose their singularity.

Despite these concerns, by the late eighteenth century, numeracy had become a commonly used skill and in the nineteenth century 'what was counted was what counted' (Cohen, 1999, p.207), generating a world characterised by being measured in every corner of its being, a world fixated on numbers (Hacking, 1990). With political arithmetic, more and more of society was quantified: mortality, suicide, crime, marriage, divorce, voting, and literacy rates, whereby numerical sociology and social statistics were created (Cohen, 1999; Desrosières, 1998; Hacking, 1990; Lazarsfeld, 1961).

In the attempt to create an 'objective social science' (Lazarsfeld, 1961, p.317) or establish a 'social physics' (Adolf & Stehr, 2018), mapping social behaviour and finding causalities were central activities. Empirical data on populations were gathered and analyses of for example marriage, death, and fertility were made with the aim of explaining patterns and connections (Desrosières, 1998). The growth of a population or the decline in marriages would be explained by the increased number of students enrolled at universities or people called into military service (Lazarsfeld, 1961). Such analyses constituted attempts to explain social dimensions and events through statistics.

Over time, the scope of measurement and what has been perceived as quantifiable has changed drastically. Measuring has gone from targeting physical objects and conditions such as temperature, to an activity that includes more and more aspects of social life and human behaviour. As Beer (2016) notes, referring to the work of Porter (1995) and Hacking (1991):

> The expansion of measurement, in the form of 'new countings' or 'new numberings', always came with some form of justification and rationale to make it seem necessary, legitimate, or important. The enthusiasm for numbers has led

> to the increasing measurement of people and to tumbling waterfalls of numbers accumulating in vast pools. (p.55)

Measurements are, in most spheres of life, unavoidable and a 'natural' consequence of societal development. Introducing new measurements and establishing them as necessary, legitimate, and important makes the call for questioning less obvious. When these legitimising effects are combined with an 'enthusiasm for numbers', there are no limits to what can be quantified, regardless of whether or not the measured object lends itself well to commensuration. One way quantification is legitimised, established as necessary, and thereby manages to escape the discussion of whether or not its object should be quantified in the first place, is by forging a strong associative link to objectivity and rationality.

Objectivity, rationality, and science

Quantitative measures are typically perceived as a superior and more scientific method than qualitative approaches because they are equated with objectivity and rationality. The fascination with quantification therefore also lies within the idea that measurement equals accountability, evidence, objectivity, and certain knowledge (Mau, 2019).

Through commensuration, the idea is that depersonalised and public forms of knowledge are produced, which are often seen as superior to more private and particular forms of knowledge (Espeland & Stevens, 1998). Moreover, quantification has been (and still is) a way of imposing order and making sense of the world (Cohen, 1999). With fast-moving change and in turbulent times, quantification has served as a way of creating order and predictability, following the notion that 'being able to measure something gives us the sense that we can control it' (Rettberg, 2014, p.62).

Although the association between quantitative measures and rationality became prevalent in the nineteenth century and continues today, it can be found 2000 years ago. As Nussbaum and Hursthouse (1984) describe, Greek writings from the fifth and early fourth century BCE indicate that commensuration, measurement, and counting were linked with order, comprehension, and control. In contrast, incommensurability was paired with anxiety and irrationality. Further, what was measurable was by nature *good*, whereas things lacking a measure were *bad,* adding a normative or ethical dimension. For example, Plato argued that ethical values and emotions could (and should) be quantified, creating

an ethical 'science of measurement' which would free us from ethical and emotional pain, uncertainties, and confusion (Nussbaum & Hursthouse, 1984, p.55). Also on other occasions, Plato was concerned with numbering evasive phenomena. In Book IX in the *Republic*, Plato concludes, through a Socratic dialogue, that the tyrant (the unjust) lives 729 times less pleasantly than the king (the just). This number is perhaps not meant to be taken too seriously, but it works to emphasise the great difference between the unjust and the just in terms of happiness, pleasure, and pain.

The weight of numbers continued through time. According to William Petty (1623-1687), who coined the term 'political arithmetic', measurements automatically imply certainty and rational thinking based on the argument that the best empirical facts about society are numerical facts (Cohen, 1999; Lazarsfeld, 1961). Similarly, John Graunt believed that a rational understanding of life was gained through numbers arranged in tables (hence the recordings of deaths) (Gregory, 2013). A couple of centuries later, psychometrician, statistician, and psychologist (to name a few of his attributions) Francis Galton (1822-1911), claimed that anything can be measured and that measurement is the 'primary criterion of a scientific study' (Gould, 1996, p.107). Likewise, Lord Kelvin (1891) stated:

> I often say that when you can measure what you are speaking about and express it in numbers you know something about it; but if you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind (cited in Crosby, 1997, p.225).

The belief that numerical measurements were methodologically superior to qualitative approaches became widespread in the nineteenth century and continues today. In a student textbook on business research methods, Cooper and Schindler (2014) wrote:

> The goal of measurement – indeed, the goal of "assigning numbers to empirical events in compliance with a set of rules" – is to provide the highest-quality, lowest-error data for testing hypotheses, estimation or prediction, or description (p.248).

Certainly, there is a strong discourse about the value and superiority of quantitative measurements.

Objectivity as a scientific goal came into being in the same period as quantitative measures became an established part of society. The concept of objectivity

emerged in the mid-nineteenth century as an epistemological goal, a scientific ideal, and a set of practices (Daston & Galison, 2010). Prior to objectivity, the virtue of 'truth to nature' (p.27) prevailed. The goal of this approach was to discover, reveal, and illustrate reality ('what truly is') by observing, analysing, distinguishing, selecting, and remembering. The scientist's role, knowledge, and experience were considered helpful and important rather than presenting a form of bias and therefore something to be reduced or eliminated (Daston & Galison, 2010).

With the emergence of objectivity, scientists feared a new knowledge obstacle: their own scientific selves. Objectivity meant the opposite of subjectivity, and therefore presupposed the suppression of the self, the negation of subjectivity Being objective appealed to self-restraint, self-discipline, self-control, and self-imposed selflessness, since 'objectivity is blind sight, seeing without inference, interpretation, or intelligence' (Daston & Galison, 2010, p.17). The biggest threat to objectivity, the epistemological danger, was the subjective self (Daston & Galison, 2010). The subjective self was thought to impose preconceptions on data, where the goal was to create an 'unclouded mirror of the world' (Daston & Galison, 2010, p.203).

There are ethical and normative connotations to objectivity. In the pursuit of scientific knowledge, certain characteristics were deemed important and encouraged. A certain *type* of scientist became the regulative ideal. On behalf of 'the common good', statistician Karl Pearson (1857-1936) directly encouraged people to suspend or repress their feelings and emotions (Daston & Galison, 2010, p.196). The model scientist was deemed to be able to 'self-eliminate', reduce 'his [sic]' subjectivity to a degree where 'his [sic]' conclusions and assessments became universally true, not contingent on the scientist's approach (Daston & Galison, 2010, p.196). Self-elimination became an imperative for scientific objectivity, meaning that 'the battle of the will against itself' (Daston & Galison, 2010, p.210) was always present in the attainment of knowledge.

Quantitative measurements became one of the most important ways of achieving objectivity, alongside photography, indicating that objectivity is about separating the researcher from the observed and measured, using instruments independent of the researcher to 'capture' reality. These techniques are supposedly free from any subjective assessment or bias, which in many cases were, and still are, considered as challenging and in conflict with professionalism and objectivity (Cohen, 1999; Daston & Galison, 2010; Porter, 1995; Rettberg, 2014). Rettberg (2014) argues that measuring leads to what we think is the objective truth, and

thereby a way of avoiding the feared bias. The concern and strive for quantification are therefore connected to and conditioned by the striving for objectivity. Objective knowledge, when perceived as the prerequisite for obtaining true knowledge, becomes the goal in science.

According to Porter (1995) the pursuit of objectivity is a pursuit of a universal language, since quantification as communication is suitable for travelling across borders. The regularity of mathematics, the rules for collecting and handling numbers are almost universal, allowing the disciplines to be practised in uniform ways across the world. Knowledge is thus produced independent from the individuals employing the numbers. In relation to this, Porter (1995) stresses how quantification works as a 'technology of distance' (p.ix), a way of making decisions seem impersonal. Quantification replaces personal judgement and allows decisions to be made 'without [decision makers] seeming to decide' (Porter, 1995, p.8), ultimately providing a way of distancing decision makers from the measured. By making numbers appear boring and technical, they appear to be beyond manipulation and human influence (Porter, 2012).

The relationship between quantification and objectivity makes them seem inseparable and each other's prerequisite or product: The idea is, that through quantification, subjectivity is contained and objective knowledge is produced. In turn, to obtain objective knowledge, a true science, quantitative measures are the most suitable tools. However, both quantification and objectivity have normative connotations and are value-laden undertakings.

The multiple meanings of 'norm' and 'normal'

Quantifying phenomena has served to (objectively) inform strategic administration of populations, govern, compare, establish statistical and behavioural norms, and ultimately identify deviations. Likewise, objectivity is a value-laden concept, an ideal or scientific norm that is put in opposition to subjectivity. It is therefore relevant to examine more closely the concept of the 'norm'.

According to Collins Dictionary, norms can be viewed as 'established standard[s] of behaviour shared by members of a social group to which each member is expected to conform' or 'ways of behaving that are considered normal in a particular society' (Collins Dictionary, 2023, n.p.). These definitions imply that norms are socially constructed. 'Established standard[s]' signals that the standards or norms are things that come into being, that *become* established. The particular

standard is not a given, but one possibility from a field of possibilities. The ways of behaving are 'considered' to be normal in a 'particular society', thereby indicating that the behaviours are not universally 'normal'. Furthermore, norms are 'shared by members of a social group', meaning that everyone in this particular social group is aware of the standards, and implicitly that people from other social groups might not consider this type of behaviour the norm. Most importantly, a standard is something to which 'each member is expected to conform', suggesting that norms contain expectations to behaviour and the power to govern conduct. Norms are therefore closely tied to expectations of conformity.

Another definition of a norm is: 'a principle of right action binding upon the members of a group and serving to guide, control, or regulate proper and acceptable behaviour' (Merriam-Webster, 2023, n.p.). This definition emphasises expectations of conformity even more. Norms operate by standardising and governing behaviour. 'Proper' and 'acceptable' do not have universally agreed upon meaning; they mean different things to different people, which is why norms can here be seen as necessary in order to cultivate this particular wanted behaviour. In other words, norms are necessary constructions to promote 'proper' and 'appropriate' behaviour, ensuring that this behaviour relies less on subjective interpretation, and rather on standards within a particular social group.

The term 'normal' has had different meanings in different contexts and periods. According to Rose (2008), it is a 'peculiar little term that condenses, in one word, ideas of the normal, the average, the statistical mean, the desirable, the healthy' (p.449). In the biological and medical domains where it evolved, 'normal' was used to describe the 'normal state' of an organism. The normal state was the healthy state, the typical and ordinary (Hacking, 1990). It was thus considered to be the opposite of the pathological. Our knowledge of normality is therefore partly derived from an interest in abnormality (Rose, 2008).

In geometry, 'normal' takes its meaning from the Latin 'norma', which means T-square, and refers to a vector perpendicular to a surface. Here, normal (synonymous with orthogonal) is descriptive: a line may be normal/orthogonal or not. However, an evaluative or normative dimension exists as well. An angle can be right (correct), but also right (normative) as in a good one. Orthodontists straighten crooked teeth, making them right, that is, even and aligned. 'Normal' can then be used to describe both how something *is* and how something *ought* to be (Hacking, 1990).

From biology and geometry, the term 'normal' moved into 'the sphere of almost everything: people, behaviour, states of affairs, diplomatic relations, molecules'

(Hacking, 1990, p.160), all of which could be categorised as normal, abnormal, or ranging on a scale between these two. In the societal sphere, normal ceased to solely mean the ordinary, and became the state to strive for. Normal became synonymous with good health and the ideal condition. This prompted the question whether a particular kind of behaviour, for example in a child, was normal. Normal can thus be understood to mean what is right, the status quo, where deviations from the norm are considered pathological. Alternatively, normal can connote the average on which something can be improved, where excellence is at one extreme of the normal distribution (Hacking, 1990). A norm can be a type of rule, a standard, or what is most common. Norms are therefore both descriptive and evaluative or even moral, which is why 'the benign and sterile-sounding word "normal" has become one of the most powerful ideological tools of the twentieth century' (Hacking, 1990, p.169).

One area in which the ideology of normality is highly influential and consequential is the sphere of psychological testing.

Social measures and their normative implications

A historical account of psychological testing

Until approximately 1850, psychology was a non-experimental branch of philosophy. The idea of quantifying human behaviour would have been dismissed (Rani, 2004). Kant, for instance, found it useless to try and measure human behaviour and reactions, since he did not consider these either observable or directly measurable (Rani, 2004). However, by the nineteenth century psychology had been impacted by approaches available to the biological sciences, to such a degree that psychologists started to strive for it to become an independent science itself. Suffering from 'physics envy' (Gould, 1996), the desire to be comparable to the natural sciences (Desrosières in Bruno, Jany-Catrice & Touchelay, 2016), and the wish to gain prestige (Kline, 1988), researchers sought to legitimise social science and psychology as 'real' sciences. Psychology formed alliances with biology, adopted the same methods and became 'measurement-conscious' (Rani, 2004, p.92).

Prevalent understandings in psychology can be exemplified by philosopher John Stuart Mill's (1806-1873) assumptions, here summarised by David Hamilton in the following three points:

> 1. The social and natural sciences have identical aims, namely, the discovery of general laws that serve for explanation and prediction, 2. The social and natural sciences are methodologically identical, 3. The social sciences are merely more complex than the natural sciences (cited in Lincoln & Guba, 1985, p.20).

This way, experimental psychology has adopted what is sometimes referred to as 'the scientific method', which, according to Kline (1988), is based upon three elements. First, rigorous, replicable, and precise observations and quantifications must be conducted under controlled conditions (original data must be quantified, hence the use of psychological tests). Second, the constructs or methods used to measure and observe must be clearly defined, agreed upon, and operationalised. Third, the method should be based on testing hypotheses which are stated in a refutable form. This stems from the logical positivist approach, more specifically Karl Popper's falsification principle, claiming that if theories are not refutable they should be discarded (Kline, 1988).

Supporting Gould's point about a dominant envy of physics, Kline likewise argues that the rationale for psychology to adopt the scientific method is based on the desire to achieve the same progress made by the natural sciences. A major goal was to establish predictive knowledge about the world. Importantly, in this argument there lies an implicit assumption that such a scientific method is the correct and only way to true knowledge.

The aim to achieve prediction and control brought with it a desire to establish causality, since causes were seen as the key to identifying predictions and potentially obtaining control. Knowledge about causes had great political power by assisting those who wished to predict and control society, supporting a deterministic scientific understanding (Lincoln & Guba, 1985). With this development followed an expansion of measurement techniques to cover all aspects of human behaviour. For example, in the attempt to determine the 'ideal average man', statistician Adolphe Quetelet (1796-1874) subjected physical attributes to quantitative analysis. According to Quetelet, the 'average man' was formed in the Creator's goal: 'perfection' (Desrosières, 1998, p.78). Later, Quetelet extended his measurements to entail suicide, marriage, and crime, numerically determining the normal, the average person. (Adolf & Stehr, 2018). With a concern for the normal distribution, Quetelet became preoccupied with

deviant behaviour. Thus, learning about the normal, the average, is closely related to knowledge about the abnormal, the pathological or irregular.

About the same time as Quetelet, Francis Galton likewise wanted to measure individual differences in physical and psychological characteristics. He singled out 'human ability' as a dimension of study and laid the grounds for psychometrics (Rani, 2004, p.95). Galton created the term 'mental test' and provided basis and direction for the study of human characteristics. Although he did not develop this, he suggested using questionnaires for measuring mental traits (Butcher, 2010).

Accordingly, the orientation shifted from a focus on external objects and phenomena to the interior space (Rose, 2008). Both psychologists and natural scientists were interested in measuring the interior realm. Gustav Theodor Fechner (1801-1887) wanted to study man's 'inner world' by applying the same methods as the natural sciences: physics, chemistry, and biology and 'psychophysics' was born (Rani, 2004, p.93).

Apart from the research stream concerned with measuring individual differences, clinical studies in medicine and psychiatry also provided the impetus for the expansion of psychological measurement. In the beginning of the twentieth century, achievement tests such as Binet's intelligence scales were developed alongside arithmetic, spelling, and language tests. During this period, Carl Jung (1907) studied word associations in order to evaluate a person's thought processes and personality (Butcher, 2010). This approach sparked enthusiasm and a group of followers or 'converts' was formed, as Rani (2004) describes them. An explosion of different tests occurred: standardised tests for educational purposes, military evaluation tests, group intelligence tests, and personality questionnaires.

Psychological tests have initially been used to identify maladjustments or mental deficiencies, and served as an expertise on individual differentiation, a technology of individualisation. Systems in need of individual administration or distribution resorted to psychological tests, based on the notion that these provided judgements what were objective, neutral, and effective (Rose, 2008). Through quantitative psychological measurements, scores could be attached to individuals, which would render the invisible visible, calculable, and manageable, permitting scrutiny and enabling judgement (Espeland & Stevens, 2008; Rose, 2008).

In response to the growing number of psychological tests, some scholars have questioned the basic assumption that psychological phenomena can be measured. Among others, Kline (1988, 1998) points to the problematic use of scientific methods in psychology. Kline (1988) argues that the scientific criteria, as

described previously, are simply not applicable to psychology since the objects of study here (personality, emotions, thoughts) are fundamentally different from the objects of study in the natural sciences for which these methods were developed. Transferring these criteria to a field like psychology is therefore problematic: Kline states that while intelligence can be measurable and capable of objective study, our attitudes, thoughts, and emotions are 'beyond measurement and … therefore, can never become scientific' (Kline, 1998, p.24). In other words, the concepts and phenomena studied in psychology are not suited for public observation and measurement (a necessity for the scientific method).

Kline (1988) further problematises how the wish to live up to the scientific demand of precision has led psychologists to choose variables merely because they are deemed measurable, not because of theoretically sound or meaningful reasons. In order to achieve precision, the number of questions that can be asked is limited; psychology formulates questions that can be answered, involving variables that can be measured. To explain his point, Kline (1988) refers to psychological handbooks of mental measurements containing for example an 'athletic motivation inventory', 'life goals evaluation schedule', and 'consumer competences test' (p.22), at the cost of, according to Kline, more meaningful studies of 'human characteristics and feelings which are important to most people' (p.22), since these are immeasurable. Ultimately, the attempt to meet certain scientific criteria comes to regulate the questions asked and how they are sought answered.

To support his statement about the unsuitable use of scientific criteria in psychology, Kline comments on arguments presented by psychologist Raymond B. Cattell (1905-1998), who advocated for scientific psychology. Cattell argues, that we apply the same methods to studies of personality as we would to studies of the mechanisms of a watch. According to Cattell the goal is always objective insight: prediction, control, and scientific laws. Kline contests the comparison made between a personality and a watch and stresses that while no one doubts what a watch is, we cannot all agree on what personality is. Personality, he argues, is a construct, which cannot exist independent of the mind that conceives it. The mechanisms are part of the conception, which is why they also have no existence beyond it. As a result, according to Kline, psychological tests do not necessarily measure what they intend or claim to measure since the very definition of certain psychological traits can be hard to agree upon. Therefore, what is measured is not self-evident (Kline, 1998). Following this criticism, psychometricians must assume that the measured traits, characteristics and the like are in fact measurable

and consisting of a quantitative structure, to which Kline responds that no such evidence is present (Kline, 1998). Even though Kline argues that psychometrics is tackling some important issues of great practical importance like selection processes, he claims that for more subtle topics such as love and other emotional matters, the contribution of psychometric tests is limited, leading him to conclude that 'psychometrics can answer some concrete applied questions but beyond this, it is defeated' (Kline, 1988, p.63), undoubtedly setting the stage for a heated discussion with scholars and psychometricians convinced otherwise.

Numbers guiding norms

Measurement activities are normative in that they statistically establish the normal levels of behaviour and turn our attention to numerical values that ought to be optimised. In relation to this, it is relevant to distinguish between different types of measurement activities.

Stein (2016) argues for the difference between physical 'facts' and psychological 'facts', which he terms 'normative facts' (p.99). Physical facts are in themselves free of normativity. For instance, a thermometer tells us the temperature, a fact free of value. Only when put in context is value added to the number. A temperature of 42 Celsius is very unhealthy for a human being, but appropriate for a hot meal.

On the other hand, social measures are value-laden even when they are not presented in context. 'Facts' in a test are either decidedly right or wrong or at least contain an evaluative scope. Unlike physical facts, this evaluative dimension is embedded in the test, not produced by context (Stein, 2016). This is important to note, since social measures are otherwise treated as factual and descriptive, simply telling us how things are, and not as evaluative.

Along the same lines, in response to the neutrality and objectivity often-attributed to social measures, Mau (2019) argues that such instruments represent 'specific orders of worth' (p.160), which are based on what can and should be measured. We might expect measurements of the social world to be neutral, objective, accurate, and rational depictions. However, they contribute to the 'establishment of the normative order' (Mau, 2019, p.160) by selecting, weighting, and connecting information in particular ways. Indeed, measurements do not just indicate value, they assign it. Through data, quantitatively perceiving worth is made possible. Things or people can be assigned their value which can then be tracked and improved. When worth is assigned, adaptability and performance

improvement are incentivised and encouraged (Mau, 2019). Orders of worth offer ways or justifications for evaluating things in a particular way, and guide our attention by telling us which activities or qualities have a high value and which do not. In this way, normative orders or principles are established (Mau, 2019, p.11). For example, by formulating and rewarding measurable key performance indicators (KPIs), certain accomplishments and activities appear as important and valuable, leaving others ignored or of less value.

The same principle applies to personality tests. The selection of characteristics or competencies to measure means that these are assigned worth. Espeland and Stevens (1998) argue that commensuration contains preconceptions of what is relevant and valuable, and thus also renders things *ir*relevant. In other words, through commensuration activities, we are told what to look at and how, systematically excluding alternative perspectives (Islam, 2022; Mau, 2019). Data institutionalises these perspectives, which in turn influences how we evaluate everything from good education to what types of performance or leadership that 'count'. Supporting this, Rose (2008) argues that numbers or tests promote certain values and norms by relying on a central tendency that people either fit or do not fit. Since 'such numbers have great power, and embody the authority of objectivity within themselves' (Rose, 2008, p.451), their use has the potential to regulate, manage, and guide people.

With numbers being associated with truth and objectivity (Cohen, 1999; Porter, 2005), the use of measurements can create certain realities, advance particular agendas, guide behaviour, and this way work as a means of control and manipulation. As Espeland and Sauder (2007) and Espeland and Stevens (2008) argue, measurements are *reactive,* in that they prompt reactions and cause people to change behaviour. In other words, numbers can guide norms and incentivise certain behaviour (Berman & Hirschman, 2018). By measuring, people's behaviour is nudged to fit the numbers (Adolf & Stehr, 2018). In this view, measurements are a source of power that offers opportunities to create and reinforce norms and realities. When attaching, sometimes uncritically, great value and objectivity to numbers and quantitative assessment tools, the results are not questioned, and their truthfulness therefore becomes a purportedly objective guideline. As Porter (1995) so eloquently says: 'Norms based on averages advertise a beguiling independence of human choice that enhances their credibility' (p.78). Statistical norms appear free of subjectivity, bias, and choice, increasing the norms' status as objective and trustworthy.

Numerical measurement of this kind can work as a powerful tool to advance what are deemed to be norms and produce disciplinary effects. Porter (1995) and Rose (2008) argue that when objectivity is associated with quantitative measurements, this becomes a way of legitimising and extending power. In other words, through such measurements, individuals can be manipulated and managed (Porter, 1995; Rose, 2008).

Quantitative measures can thus be used to simplify, classify, compare, and evaluate (Espeland & Stevens 2008). Assessment measures contain a statistical normal which can easily be conflated with a behavioural or moral normal (Espeland & Sauder 2007). Measurements and numbers express different ideas of normalcy which some critics argue creates 'an oppressive language of normality and abnormality' (Hacking, 1990, p.1; Porter, 1995, p.77). Through classification schemes, rankings, and the act of scaling, individuals are hierarchised and their place (in the normative sample) is made known (Townley, 1993). Quantifying individuals according to a scale can work as a powerful way of creating norms, direction, codes for legitimate and preferable behaviour, and ultimately work as a normalising process. As Hacking (1990) argues, data about averages promote an idea of normal people and a quest for modifying undesirable behaviour (Hacking, 1990). People are conceived normal, when they conform to fit the average, the central tendency, since the extremes are considered pathological.

As a result, most people try to be normal, which in turn affects what is then perceived as normal. Moreover, through measurements, people become comparable with other people and targets emerge that one either ought or wish to reach (Mau, 2019). In short, the emergence of political arithmetic and a concern with normality meant that: 'The cardinal concept of the psychology of the Enlightenment had been, simply, human nature. By the end of the nineteenth century, it was being replaced by something different: normal people' (Hacking, 1990, p.1).

Measurements in organisations

Measurements and psychological tests have infiltrated organisational life in a number of different ways. Commensuration has historically been seen as fundamental to management, steering decisions on everything from 'welfare to warfare' (Espeland & Stevens, 2008, p.324). This has to do with commensuration being associated with rationality, leading to a preoccupation with continuous and

most importantly, measurable, improvement. The emergence of a need to document and track goals, competencies, and behaviour, more specifically, the rise of management by objectives, balanced scorecards, 360° feedback, Key Performance Indicators, and SMART-goals (M for Measurable), shows a conviction that these measurements are reliable and beneficial when tracking and assessing behaviour, performance, and productivity. In other words, the widespread use of quantitative assessment tools is based on the belief that quantification is the primary and preferable technology of performance management (Espeland & Stevens, 2008). In this view, tools such as the 360°, 'render individuals observable, measurable, and quantifiable' (Townley, 1993, p.529). This and other appraisals share the presumption that measurement of performance will lead to improvement, resting on the notion that: 'Unless something can be measured, it cannot be improved' (Kelly, 2007 cited in Moore & Robinson, 2015, p.7).

In the late nineteenth and early twentieth century, the economic notion of efficiency was the undisputed rationality behind Fordistic and Tayloristic disciplines. With the Efficiency Movement, performance and behaviour were attempted measured in the name of efficiency and utility maximisation. Over time, this rationale has extended and measures of work now exceed the engineering of processes, labour productivity, and workflows. Today, different technologies and wearables aimed at tracking employees' productivity are all used in the name of efficiency. Self-tracking devices are introduced to encourage employees to adjust their behaviour and improve their productivity. More aspects of employees' lives, e.g. their health and how they manage their time, have been made subject to monitoring, measurement, and consequently, control (Moore & Robinson, 2015).

Prior to 1950, the goal of using personality tests was to identify maladjustments or the abnormal. Using tests originally developed for the military in World War I to identify 'unstable soldiers' or screen 'at-risk recruits' (Gibby & Zickar, 2008, p.166), managers in some organisations were preoccupied with 'rooting out undesirable and unstable workers' (p.167). This was further legitimised due to psychologists' estimation that '80% of problem employees had a "quirk or unusual feature" in their personality' (Humm 1943 cited in Gibby & Zickar, 2008, p.167). The belief was that by identifying and eliminating these maladjustments, productivity would increase.

The Thurstone Personality Schedule used in 1936 to investigate adjustment difficulties of female teachers showed that: 'One third of the women teachers are

definitely maladjusted, and one sixth need psychiatric advice … Only one fifth can be classified as well-adjusted' (Gibby & Zickar, 2008, p.170). This example shows how a test built on statistical norms about maladjustment simultaneously promotes such norms by identifying those who do not fit the criteria for a 'well-adjusted' teacher.

In the educational system, tests have generally served as a means of categorisation. Ability tests have been used to sort and segregate students, based on either skills or level of talent (Danziger, 1990; Porter, 1995). Intelligence tests and statistical analysis created a scientific basis for these decisions, and provided a legitimising rationale allowing individuals to be classified, a necessity in bureaucratic organisations, including schools and large workplaces (Danziger, 1990).

Porter (1995) raises concerns about the implementation of standardised tests in the educational system and argues that while these tests appear to provide 'impersonal objectivity' (p.210), in fact they include unfairness, racial bias and function at the expense of teachers' expertise. Stein (2016) adds that test use in education risks reducing social efficiency to economic efficiency, and learning to test score gains. Efficiency trumps justice, which eventually, paradoxically, leads to inefficiency. Exemplified here in the educational system, tests provide rationales for, at times, unjust categorising and they risk reducing concepts, such as learning, to a simple score.

Measuring and conceptualising leadership

Established in the organisational sphere, quantification naturally also infiltrated leadership (development). How leadership concepts have evolved is closely tied to attempts of measuring such concepts quantitatively, which is why both the development of leadership concepts and measures are intertwined (here and in practice). Interestingly, the more immeasurable the concept of leadership appears to become, the more tools to quantitatively measure it are developed, supporting the notion that what we choose to measure, is what we wish to make certain (Cohen, 1999). With leadership becoming more and more intangible, at times spiritual, all in all immeasurable, the greater the need becomes for objective guidelines, formulas, and behavioural prescriptions.

The use of quantitative measures in leadership research and leadership development can be traced to the Ohio State Studies around 1945 (Bass, 1990). At this point in history there was little measurement practice in management and organisation research. The Ohio State Studies identified two key concepts called

consideration and *initiating structure* and introduced operational measures of these (Schriesheim & Kerr, 1974). This eventually led to the construction of the 'Leadership Opinion Questionnaire', 'Leadership Behavior Description Questionnaire', and the 'Supervisory Behavior Description Questionnaire', collectively comprising the 'Ohio State Leadership Scales' (Schriesheim and Bird 1979; Schriesheim and Kerr 1974). These questionnaires are based on both self-report and others' rating. The Ohio State Studies thus laid the groundwork for so-called leadership measures. Besides offering leadership scales and operational measures, these studies contributed to a concern with what leaders *should* do, defined by Barrow (1977) as 'normative leadership approaches' (cited in Bryman, 1986, p.75). These approaches are built on moral principles and norms for how leaders ought to act.

Many leadership (development) scholars, both past and present, operate within a functionalist discourse (Mabey, 2013). Based on positivist assumptions, this approach is concerned with trait development, measurable, individual improvements (Mabey, 2013), and skill acquisition (Lord & Hall, 2005). Within this discourse, leadership development is concerned with identifying causal relationships, 'disassembling and reassembling the leader' (Barker, 2001, p.484), and building competencies, overall reflecting a general emphasis on instrumentality (Lord & Hall, 2005).

Leadership measures therefore represent a certain research tradition, particular ontological and epistemological assumptions. Developing leadership measures in an attempt to quantitatively measure leadership follows a belief that the concept of leadership manifests itself and take some sort of form that is then possible to capture through the use of numbers, independently from the social context. Measuring leadership suggests that there are clear and measurable dimensions that together constitute a person's abilities as a leader. Attributes are perceived as quantifiable in a one-dimensional way. Assumptions about measurable personality traits, their predictive strength, and knowledge about what constitutes successful leadership are all dominant (e.g. Barbuto & Wheeler, 2006; Walumbwa et al., 2008). The rationale is, that by asking the right questions one can gather truthful information about something external. It is therefore apparent that the underlying ontology lends itself to the positivist tradition, where objective knowledge and explanations are made possible using certain scientific methods that build on notions such as validity, reliability, and replicability. Following a positivist line of reasoning, leadership is something that exists 'out there', about which one needs to collect and systematise data and then make predictions about

how it truly operate in reality. In sum, the measures are assumed to be able to provide an objective picture of an individual as a person or leader, which is then the presumed prerequisite for improvement and development.

The way leadership understandings have developed has implications for how one tries to measure and assess leadership. One essential change in common leadership understandings is related to the pivotal shift in leadership studies in the 1970s, when a theoretical distinction was made between leaders and managers. Zaleznik (1977) asserts, in unambiguous, almost poetic terms, that leaders differ greatly from managers, distinguishing leaders by attributing to them extraordinary, almost superhuman, qualities – all that managers are not. In this light, contemporary leadership constructs are influenced by Thomas Carlyle's (1795-1881) notion of the Great Man (Spector, 2016; Spoelstra, 2018). According to Carlyle, Great Men are so-called 'light-fountains', whose light, the 'gift of Heaven', a 'force direct out of God's own hands' enlighten 'the darkness of the world' (Carlyle, 1840, pp.4; 16). This perspective brings a leader-centric focus on the impacts of unique, hero-like individuals with superior powers or qualities, while at the same time acknowledging the critical role of followers. Great Men, heroes, and their followers are deemed to be the foundation of society, of the hierarchy or 'hero-archy' (Carlyle, 1840, p.15), as Carlyle points out – drawing attention to the Greek *hierarchēs* (*hierós*: sacred, holy + *archēs, archos*: ruler, leader, prince).

Half a century later, Max Weber, whose work was later republished (1968, 1978) developed the concept of 'charismatic authority'. According to Weber, charismatic leaders are extraordinary, as they have, or are perceived to have the 'god-like strength of the hero' (1968, p.24) and 'exceptional powers' (1978, p.241). Their leadership is based on 'inspiration', 'divine judgments', and 'revelations' (Weber, 1978, p.243). Weber explains: 'Charismatic domination means a rejection of all ties to any external order in favor of the exclusive glorification of the genuine mentality of the prophet and hero' (1968, p.24). It is precisely the rejection of ties, the break from bureaucracy that enables devotion for the 'unheard-of', the 'strange to all rule' (p.23). This way, charismatic leadership stands in opposition to the ordinary, formal, traditional rules and laws.

Influences from theorists such as Carlyle and Weber have led to what some scholars characterise as ideological, leader-centric, and romanticised leadership (e.g. Alvesson & Kärreman, 2016). The antagonistic relationship between the ordinary and the routine and the extraordinary and 'unheard-of' means that the object of leadership studies has changed.

Contemporary leadership constructs bear many resemblances to the concepts of charismatic leadership and Great Men ideas. For example, by introducing the concept of transformational leadership as an alternative to transactional leadership, Bass (1985) contributed to the cementation of the distinction between management and leadership. Where transactional leadership is based on the exchange of work and rewards, transformational leadership is about instilling motivation, confidence, and consciousness so that our performance exceeds the expected and we transcend our self-interests to consider the greater good of the organisation (Bass, 1985). Bass directly links charisma to the transformational leader on several accounts. Charisma is what separates the 'ordinary manager' from the 'true leader' (Bass, 1985, p.34), who, as a result is met with stronger feelings of either love or hate. Ordinary managers do not have the power to instil these feelings. Furthermore, transactional leaders are working from within a culture, suggesting that this individual is part of the system, the routine. Transformational leaders on the other hand, are outside the system. They have an outside-in perspective on the organisation, allowing them to evaluate the culture, form a vision, and start implementing change (Bass & Avolio, 1993).

What is of particular interest here is that the more that incorporeal, hardly measurable, qualities have historically been assigned to leadership, such as charisma, authenticity, and spirituality, the more measures have been developed with the purpose of measuring and capturing exactly these ephemeral qualities (e.g. the Multifactor Leadership Questionnaire and the Authentic Leadership Questionnaire). This appears to be based on the conviction shared by Authentic Leadership Questionnaire developers Walumbwa, Avolio, Gardner et al. (2008) who argue: 'Simply expecting leaders to be more authentic and to demonstrate integrity will be ineffective if tools for measuring these aspects of leadership are lacking' (p.90). This statement suggests that certain types of behaviour can only be cultivated by means of quantitative measurement tools. This further implies that any kind of leadership construct must be tied to a measure in order to have any real effect.

Besides a perceived need to capture and cultivate immeasurable qualities in individuals recognised as leaders, particular expectations of leaders likewise pave the way for more leadership development and more quantitative assessment tools targeting the inner selves. Organisational trends and management fads likewise drive leadership skills training. The widespread idea that 'for organizations to survive and succeed through such demanding conditions, exceptional leadership

is needed at all levels' (Dalakoura, 2010, p.434), means that leadership development programmes are high on organisations' list of priorities.

Organisations also call for personal and emotional investments, meaning that it is the leaders' inner selves that are targeted. For instance, when arguing for more spiritual leadership, Fry (2003) emphasises the need for leaders to practice different spiritual rituals: to know oneself, be trusting, and 'maintain a spiritual practice' by, for example, spending time in nature, praying, meditating, reading 'inspirational literature', practising yoga, or writing a journal (p.704). By committing oneself and fostering these personal dynamics, the conviction is that followers' intrinsic motivation, joy, peace, and serenity with increase. In sum, 'leadership is not a rational endeavour; it is a deeply emotional and psychological one' (Wood & Petriglieri, 2004, p.217).

The demand for more and more qualities in and personal investment from a leader contributes to the pronounced need for leadership development and assessments that cultivate this. An assumption exists (for some organisations and certainly some test developers) that leadership is the solution to almost any organisational challenge. Therefore, a vast variety of leadership development programmes have been launched at consultancy agencies and within organisations. These have the purpose of assessing and developing the participants' potential.

Through a review of the literature on leadership (development), it appears that the call for leadership measures is multidimensional. While acknowledging that others may exist, several factors are identified here. First, strong development discourses infiltrate organisations, influencing leaders and employees. In many organisations there is a prominent concern with self-development, improvement, and achievement, which encourages leaders to request the tools themselves. Second, leadership measures are commercialised and marketed as a necessity and their implementation in organisations can therefore also be interpreted as a type of legitimatisation strategy. Mimicking others, organisations resort to quantitative assessments in order to support and perhaps account for decision making and ensure valid and evidence-based tools for development, avoiding subjective assessments or the always dreaded bias. Third, the act of measuring leadership is 'an attractive concept that seems to promise precise, scientifically valid 'proof' that a person has (or does not have) appropriate qualities' (Lashway, 1998, pp.2–3). Accordingly, the appeal of leadership measures partly stems from the pursuit of some sort of reliable proof upon which one can rely for taking actions. Test proponents and test sales representatives argue for this proof by describing tests as 'eliminating any bias' (Global HR Research, 2020, n.p.), 'free from human

judgement', by providing 'objective comparison' (Greenthumbs, 2022, n.p.), and an 'objective framework' (Thomas, 2022, n.p.), all in all, making the process 'consistent and fair' (Greenthumbs, 2022, n.p.).

Lastly, the distinction between good and bad leaders means that in order to find and develop successful leaders it is necessary to identify, 'the qualities that differentiate the "best" from the "poorest" leaders' (Van Dusen, 1948, p.67). A quote from 1948, but a concern that continues today (e.g. Ivanov, McFadden & Anyu, 2021). Measurements are here perceived as advantageous in clarifying this differentiation and establish what constitutes a good leader. Following this rationale, measurements and 'scientifically valid proof' can enable one to develop the necessary identified components of leadership in more people. As Lashway (1998) puts it: 'Because of such questions, there is always a strong market for instruments that promise valid and reliable measurement of leadership qualities' (p.2). As a result, tests have become increasingly popular, laying the ground for more books on the subject, adding again to the acceptance of 'psychological testing as an integral element of society' (Borsboom, 2005, p.1).

Endless tests and tools have been and are still being developed, more or less aimed at leaders. Despite clear and seemingly convincing claims and promises, it is less clear *what* these tools are actually measuring.

What constitutes leadership, what is being developed and measured in leadership development programmes and questionnaires varies and is referred to in many different ways, such as: 'traits' (Shamir & Eilam, 2005), 'abilities', 'skills' (Borsboom, 2005; Mehrabani & Mohamad, 2015), 'characteristics' (Mills & Boardley, 2017; Shamir & Eilam, 2005), 'techniques', 'capacities', 'personality factors' (Borsboom, 2005), 'attitudes' (Borsboom, 2005; Mills & Boardley, 2017), 'self-attitudes' (Mills & Boardley, 2017), 'behaviours' (Nielson, 2011), 'emotions' (Avolio & Gardner, 2005), 'attributes' (Borsboom, 2005; Mehrabani & Mohamad, 2015; Shamir & Eilam, 2005), 'qualities' (Lashway, 1997; Mills & Boardley, 2017), 'components' (Avolio & Gardner, 2005; Mehrabani & Mohamad, 2015), or 'capabilities to develop (inspiration, motivation, environment of trust, communication, team work, creativity, empowerment, effectiveness, employee performance and satisfaction, and knowledge sharing)' (Mehrabani & Mohamad, 2015).

Already an ambiguous picture takes shape. What is described as being measured varies from something inherent, almost material like, 'skills', 'techniques', 'traits', or 'abilities', to something equivocal like, 'emotions' and the 'capabilities to develop …'. Most importantly, a study of the field shows that endless definitions

of leadership mean endless ways to measure it, which is why 'every test thus reflects a particular set of assumptions about leadership' (Lashway, 1997). This is supported by Kaplan (1964) who states: 'How we put the question reflects our values on the one hand, and on the other hand helps determine the answer we get' (cited in Messick, 1980, p.1021). Accordingly, test results reflect test takers' perception of their performance using the available vocabulary and set scale (Lashway, 1997). Taking this even further, one could argue that test results merely show where people have placed their mark, in other words, the items or levels of agreement they have chosen among the available options.

The countless leadership definitions and term confusion indicate that measuring leadership is perhaps not as straightforward and streamlined an activity as it is officially presented by test practitioners. What is targeted in leadership development programmes, what is measured in tests, and what assumptions are at play all vary significantly.

Leadership measures are equivocal in other ways as well. Besides producing data about averages and normal behaviour, leadership measures are prescriptive in that they measure leadership: a value-laden concept in itself (Ciulla, 2004). Ciulla (2004) argues, that 'ethics is located in the heart of leadership studies' (p.4), which is why leadership both guides and is guided by values that direct choices and actions. Most contemporary leadership concepts and theories thus have strong messages about how something should be done, providing behavioural prescriptions (Jones, 2007).

The aim of leadership measures is to encourage improvement or change. In other words, measures prompt reactions (Espeland & Sauder, 2007; Espeland & Stevens, 2008), meaning that individuals will 'alter their behavior in reaction to being evaluated, observed, or measured' (Espeland & Sauder, 2007, p.11). Often through what Espeland and Sauder (2007) call 'self-fulfilling prophecies' (2007, p.11), reactions to (social) measures confirm the expectations or predictions embedded in the measures by encouraging behaviour that conforms to these expectations or predictions. For example, in university rankings, the object of Espeland and Sauder's study, certain selected criteria are used. The rankings highlight differences in universities that might not actually be considerable. They may even constitute mere statistical noise. Nevertheless, this prompts the reader to assign more worth to the top-tier schools. A university's ranking also influences budgets and therefore resource allocation to develop the quality of the school, further establishing the advantages and disadvantages among the universities. Moreover, previous rankings impact how people answer other or future surveys

about an institution, consequently strengthening former judgements (Espeland & Sauder, 2007).

Returning to leadership measures, they both rely on and promote norms. The content of the tools is constructed according to social and cultural norms, based on decided necessary leadership components, and their logic then depends on a statistical norm. By comparing scores to a statistical norm, each test taker's result is assessed normatively. Through this process, certain levels of scores, that is, levels of normative behavioural expression, are promoted.

The test industry

How leadership measures are developed, what they target, and how, vary a great deal. A tool can be categorised as being either scientific or commercial. Academics, usually working at a university, are primarily concerned with constructing what they would claim to be a scientific instrument. The goal is to contribute to research on leadership (measures) and advance science. Examples of such scientific measures are the Authentic Leadership Questionnaire, developed by Avolio, Gardner and Walumbwa, and the Servant Leadership Questionnaire, developed by Barbuto and Wheeler. As Barbuto and Wheeler (2006) state, they assembled an 'expert panel' consisting of six leadership faculty from three universities and five leadership doctoral students from one university (Barbuto & Wheeler, 2006). The measure has thus been developed within academia and mainly promoted in academic journals. However, some scientific measures are later commercialised.

In contrast, commercial measures are developed on behalf of a commercial (often international) company. Test developers here are concerned with producing and selling a commercial product. The measures are usually developed by psychologists, statisticians and/or psychometricians, and programmers. One of the largest industry players is Hogan Assessments. Hogan started as a small start-up in 1987 founded by Joyce and Robert Hogan. In 2019 they described themselves as the 'industry leader', with the goal of 'improving the global workforce', made possible in part by having a strong 'Hogan brand' (Hogan Assessments, online, n.p.). For these types of measures, the purpose is explicitly to have an effect on the workforce, which is why branding and public exposure play significant roles. The objective is practical impact, for example through the development and improvement of today's leaders. The test developers' target group is therefore practitioners, that is, potential buyers.

Depending on the way and by whom the measures are developed, they have different purposes and implications. Different things are at stake, so to speak. Scientific measures are promoted in academic journals and their acceptance is first and foremost sought there. Commercial measures are marketed as more traditional products and services, by creating a brand communicated through appealing and persuasive websites.

Besides this distinction, leadership measures and tests can be divided into self-assessments and 360° or multi-rater/source assessments. Self-assessments are, as the term indicates, based solely on the test takers' self-report or self-perception. Multi-rater or 360° assessments are based on several evaluators.

Further, the tools target different things, depending on what test developers find is the most valuable and important to measure. As mentioned earlier, some developers claim to measure personality, while others attempt to assess behaviour or reputation. Most self-assessment tests are usually claimed by test practitioners to measure either personality or behaviour. In contrast, 360° assessments measure individuals' reputation and how others perceive them.

Finally, self-assessment tests can be either normative or ipsative. In normative tests, such as those developed by Hogan Assessments, respondents are usually presented with one statement at a time to which they indicate their level of agreement using a Likert scale. For example: 'I easily get distracted when I am working' followed by five options: 'strongly agree', 'agree', 'neutral', 'disagree', or 'strongly disagree'. These tests usually generate descriptions of the respondent's personality, preferences, or behaviour. The descriptions are based on the respondent's scores on different scales, in other words, where the respondent scores compared to the norm. The fundamental mechanism of normative tools is thus a comparison between the respondent and the norm group. The norm group is created by collecting the test scores of a 'relevant' and 'representative' group of people, enabling one to benchmark later test takers' scores. In psychometric assessments there are usually different norm groups, e.g. a global and a national one, or norm groups for different professions such as 'sales people', 'executives', and 'students'. Depending on what norm group the test practitioner considers it relevant to compare the test taker with, the scores come out differently.

In ipsative tests, respondents are forced to choose what is most and least true to them out of several statements. The respondents are not compared to a norm group, instead they are assessed in relation to themselves, their own preferences, and their own mean, so to speak (Cattell, 1944). Ipsative tests are often typology

tests, where the result generates a category, colour, or letter (combination) to which the test taker belongs, according to the test. In this way people are grouped into different types. DiSC is an example of an ipsative test. Here, respondents' 'behaviour profiles' are characterised by one letter, either D (dominance), i (influence), S (steadiness), or C (conscientiousness), or a combination of letters. The Myers Briggs Type Indicator is another example of an ipsative tool. Here, respondents fall into one of sixteen possible personality types, as indicated by four letters and their combinations.

Depending on who one consults, normative and ipsative tests have different strengths and weaknesses. Bowen, Martin and Hunt (2002) argue that ipsative tests are harder to game or fake than normative tests, whereas Hicks (1970) advises against their sole use, due to their 'extensive psychometric limitations' (p.167). Not surprisingly, those developing, selling, and representing a certain test tend to highlight their benefits. It would seem that the preference for normative or ipsative tests is a matter of more or less active choice or belief.

How tests are actually developed is complicated and not very transparent. In explaining the process, Robinson (2017) uses this figure:

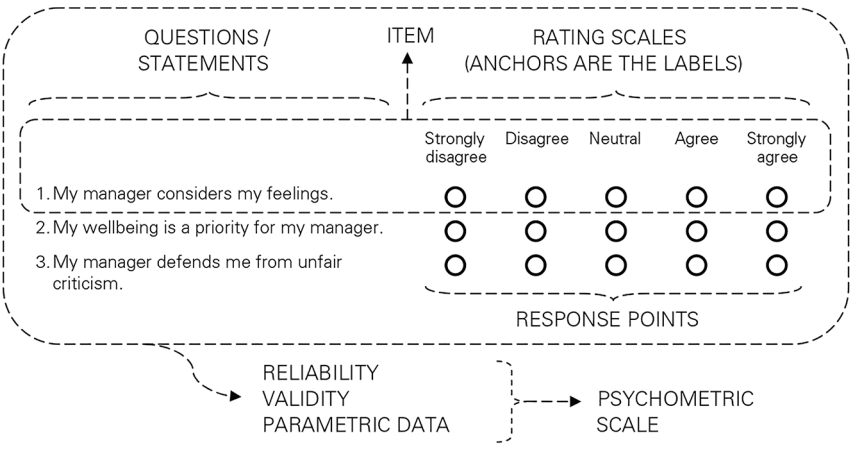


Figure 1: 'Components and characteristics of a psychometric scale' (Robinson, 2017)

The test items, whether questions or statements, reflect a 'focal variable' which the test developer wishes to measure, for example 'dominance'. A psychometric scale is then defined as 'multiple items measuring the same focal variable in a

reliable and valid manner' (Robinson, 2017, p.740). Items or scales are identified in different ways. Test developers can either use an existing scale or develop their own. For example, the Jefferson Scale of Physician Empathy is a 20-item scale 'with sound psychometric support' (Kane et al., 2007, p.83) broadly used to measure physicians' level of empathy. Alternatively, according to Robinson (2017), items can be identified through 'literature reviews, interviews with experts, and content analysis of existing data sets and resources' (p.742). For example, in the *Leadership Effectiveness Analysis: Technical Considerations Report* from 2010, the authors describe under 'Origins of items: Theory/research' that the test was constructed by 'observing leaders and attempting to identify those behaviors and practices that tended to lead to success over a wide range of leadership challenges' (LEA Technical Considerations, 2010).

According to an 'expert committee' formed by the Board on Testing and Assessment of the American National Research Council who refers to the Standards for Educational and Psychological Testing[2], one must first define the purpose of the test, that is, what it is supposed to measure. Second, 'test specifications' are made, including for instance 'how the test questions will sample from the larger construct', the number of items, the format of these, and the 'desired psychometric properties of the items' (Committee on the U.S. Naturalization Test Redesign, 2004, p.13). After this 'content framework' (p.13) is defined, a set of potential test items that meet the test specifications is created.

It is not explained in detail how test items are created or how they are assessed to meet the test specifications. The quality of the test items and the developed 'scoring rubrics' are reviewed by a 'panel of experts' (Committee on the U.S. Naturalization Test Redesign, 2004, p.13). The test items are incorporated into a questionnaire which then undergoes a 'pilot test' where the psychometric properties are evaluated, such as an item's difficulty to be understood or its bias. Items that are considered to meet the test specifications will then form the final test. Through factor analysis for example, the intercorrelation between items and the degree to which items are measuring the same variable is determined. Following this rationale, if test takers agreeing on item A also agree with items B, C, and D then these items are believed to be measuring the same variable/characteristic/trait.

[2] A joint publication of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999)

Despite the seemingly controlled process of test development, it is important to also highlight how test development can be influenced by personal ideas, beliefs, and preferences. For instance, the Myers Briggs Indicator, developed by Katharine Cook Briggs (1875–1968) and her daughter Isabel Briggs Myers (1897-1980) in the 1940s, is the result of Katharine's personal obsession with Carl Jung and his Type Theory. She reported that Jung appeared in her dreams, his conception of types gave her life new meaning, and she studied his words with reverence and commitment not unlike religious followers studying the Bible. As Emre (2018) writes: 'Jung became her "personal God", and Katharine a disciple' (pp.37-38). Emre (2018) continues that Katharine identified a problem: science was too 'impersonal and objective', lacking the data of the 'soul' (p.3). Katharine thus sought to combine science and spirituality, focusing on 'the individual' (p.3).

I include this example to draw attention to the subjective, human side of test development; Katharine Cook Briggs was driven by her passion for typification and classification of people, and her personal convictions about how one better understands the world.

Studies on test use: positioning and contribution

In the next two subsections I review previous studies on tests. The first section includes a brief account of the studies to further validate and promote tests and studies to map and explain test reactions. The latter is included since their overall aim is more closely related to that of my own study, but methodologically designed very differently. Moreover, this section provides a picture of the field I wish to challenge, by exploring the phenomenon in a methodologically contrasting way. Following this, in the second section, I review studies more closely related to mine, both in purpose and methodology. These studies represent the field I contribute to and extend.

Testing the tests

In line with the functionalist approach by which quantitative assessment tools are developed, most studies on test use are likewise quantitatively based. A group of studies is concerned with proving the benefits of tests or identifying areas of improvement (e.g. Arthur, Woehr & Graziano, 2001; Datta, 2015; Melamed & Jackson, 1995). Melamed and Jackson (1995) suggest that psychometric instruments empower individuals, ensure informed decisions, and optimise team compositions and as a result, team effectiveness and team performance.

Other studies focus on the different elements of the test and point to potential practical problems and issues people should be aware of. Ni and Hauenstein (1998) explore the relationship between item invasiveness and face validity on applicants' affective reactions towards the selection process, and the interrelationship between invasiveness, face validity, and actual job-relatedness. Brett and Atwater (2001) look into reactions to feedback and the significance of the feedback source. In contrast to the previous mentioned studies, these two are less concerned with test purpose and more with other factors that can influence test reception.

Actual test reactions have also been investigated quantitatively by putting the test to a test. Scholars here do not question the test's scientific foundation, basic assumptions, and logics, in fact they reproduce them, but they acknowledge the significance of individual reactions and beliefs. The studies below are concerned with reactions of job applicants taking tests as part of a selection process. Converse et al. (2008) set out to investigate the effect of test formats and warnings on faking on 'test-taking ease, test-taking anxiety' (Converse et al., 2008, p.167). Chan et al. (1998) quantitatively measure pre- and post-test reactions. More specifically, they measure if 'belief in tests' impacts on test performance and if test performance impacts on post-test reactions. Visser and Schaap (2017) also explore job applicants' attitudes towards tests, acknowledging that participants' perceptions can impact the results. They measure six areas: 'motivation, lack of concentration, belief in tests, comparative anxiety, external attribution, and future effects' (Visser & Schaap, 2017, p.5). The study is based on the conviction that 'attitude has a profound impact on the performance of an individual', and therefore the aim is to 'come to an understanding of individual performance levels in assessments' (Visser & Schaap, 2017, p.1). Importantly, this aim is in regard to test validation and ensuring that the results are reliable. This information is supposed to improve the quality of test assessments and their interpretations.

The studies on test reactions show that regardless of the science behind these tests and the degree of validity, recipients have certain reactions to tests and feedback. Scholars conclude that test reception depends on the character and source of feedback, degree of face validity, and the tests' usefulness/applicability (concerning time efficiency etc.). According to Kline (1998), technically, face validity has no necessary connection to true validity. However, in practice these studies show it is greatly significant: Participants' perception of face validity influences their reaction to items' invasiveness. That is to say, the more relevant the question is perceived, the less invasive it appears.

The above-mentioned research is not concerned with a qualitative investigation into the subjective, emotional experiences, involved with test taking. To optimise and validate the test, these scholars assess reactions to feedback, that is, whether participants find their test results useful and accurate. However, they do not focus on the variety of reactions, subjective experiences, and how individual test takers express these. By employing a set framework or questionnaire, only restricted answers and experiences can be found (Cicourel, 1964). Finally, scholars in this field do not question or explore the acts of measuring and quantifying psychological traits.

Most importantly, scholars in all the above-mentioned studies use the same methodological toolbox as the developers of the tests they evaluate, leading to the paradoxical activity of testing people's belief in tests. Testing a test in the way noted above involves basically using the same assumptions about the measurability of attitudes and emotions, and the value of commensuration and statistics. Thereby this investigation does not question the test activity itself or its assumptions. Consequently, the knowledge produced in these studies is limited: Test takers' responses are restricted to include only the available options. There is little or no room for nuances, ambiguities, and contradictions in emotions and attitudes despite the fact that these are characteristic of human experiences.

A critical take on quantification

In contrast to the many quantitative studies typically undertaken to further advance and improve test use, a number of more critical, qualitative studies focus on the problematic aspects of quantifying human behaviour. These represent a methodological and epistemological approach and style that resemble mine.

Scholars voicing critical or sceptical perspectives emphasise the normalising and potentially discriminatory effects of social evaluative measures, exemplified by rankings of universities and ability tests in schools (e.g. Hacking, 1990; Porter, 1995; Espeland and Sauder, 2007; Rose, 2008; Mau, 2019). The overall argument in this stream of literature is that numbers and behavioural averages promote ideas about normalcy and invite (or push) individuals to conform to fit the desirable, numerical and behavioural standards. This argument contests the positivist claim that statistics merely mirror reality.

Others criticise the assumptions in and consequences of questionnaire use. According to Cicourel (1964), forced responses restrict 'out of the box' thinking or 'problematic' perceptions and interpretations. The questions asked in the test supply respondents with clues about the questions' purpose and expected answers,

generating guided and standardised responses (Cicourel, 1964). The test offers certain options or a particular scale, limiting the possible outcomes.

Alvesson and Einola (2020) likewise problematise the use of questionnaire formats and the general tendency to quantify complex social phenomena such as leadership. The authors argue that surveys give rise to various types of ambiguities such as different interpretations of wordings, poor contextual fits, or low relevance of items. They encourage a more critical approach to survey use and methodology in general which is sensitive to ambiguity: 'Just because numbers add up and correlations are produced, does not automatically mean that knowledge obtained stands on a firm footing' (p.8).

Serving as inspiration for a number of studies critical of test use, Foucault (1991) argues that measurements and rankings classify, hierarchise, and normalise individuals. Although referring to examinations, Foucault argues that through rankings, characteristics or skills can be hierarchised, and punishments and rewards distributed. Quantifying phenomena such as leadership can be seen in this light as an attempt to rank, and thus hierarchise, normalise, discipline, and potentially punish or reward. Along the same lines, Rose (2008) argues that tests work as a device for:

> Capturing these ephemeral behaviours, the evanescent qualities and variable capacities of human beings, rendering them into thought as 'docile' objects. Test scores – tables, graphs – as immutable mobiles – enable the stabilization, accumulation of information about the subjects of testing. They enable them to be normalized, tabulated and deliberated about in the calm situation of the psychologist's office (p.450).

Further sub-arguments are for example that numbers and indicators are harder to challenge than judgements based on what seem like mere opinions, since people are more inclined to trust what they perceive as hard facts than gut feelings or opinions (Mau, 2019). According to Mau (2019), the process of turning social phenomena into quantities is a way of avoiding justification and criticism. The immunity to criticism is fuelled by the conviction that 'data never lie' (Mau, 2019, p.160), contributing to people ignoring or forgetting that numbers have meanings which can be used for political purposes and interpreted to promote particular interests. Following the conviction that data never lie, and that numbers embody authority, Rose (2008) stresses social measures' ability to manage, distribute, and 'administer' individuals (p.451).

There are a number of qualitative studies on test use in leadership development that take the broader context into account. Often from a Foucauldian perspective, these researchers explore identity regulation and disciplinary effects, acts of resistance, and the production of 'confessional cultures' (Ferry & Guthey, 2021; Gagnon & Collinson, 2014; Wilson et al., 2020). These studies have in common a focus on practices and interactions that take place around the assessment tools, acknowledging the broader context in which assessment tools operate. Moreover, the authors do not treat such measures as neutral instruments, or leadership development programmes as neutral arenas. Rather, they are interpreted as means and places for identity regulation and technologies of subjectification and normalisation.

For instance, Ferry and Guthey (2020) explore how leadership development programmes for students in universities produce and normalise what the authors term a 'confessional culture of leadership development'. Through 'inward-focused, quasi-therapeutic' practices such as icebreakers and assessment tools, participants are encouraged to alter their identities in the pursuit of e.g. authentic or transformational leadership (Ferry & Guthey, 2021). Of greater significance, the participants are also encouraged to share their insights, strengths and weaknesses, that is, their test results. Icebreakers and assessment tools which the authors describe as 'confessional technologies', produce this culture, one that normalises compliance and submission to these types of identity-reshaping technologies.

Meier and Carroll (2019) are, with inspiration from Hacking, interested in how leaders are 'made up' or 'produced' in leadership development programmes. The authors explore how leader identities are constructed and authorised, paying attention to not just the personality test itself but its setting: the role of instructors and programme participants and the interaction between them.

Wilson et al. (2020) have likewise completed a critical study on measures, specifically 360° instruments. From a Foucauldian perspective, the authors seek to uncover the assumptions and unintended effects of 360° tools. They reveal such measures' morally prescriptive standards and capacity to promote surveillance, both of which force individuals to self-discipline and conform through processes of subjectification. According to the authors, the format of 360° tools discourages critical questions, leading to what the authors argue is 'inhibiting rather than enabling the development of ethical leaders' (p.213).

Also highlighting the context of measurements, Elmes and Costello (1992) explore the significance of the venue and of social actors' role in training activities. The authors argue that consultants create an aura of status and credibility, manipulating participants to feel emotionally indebted to their organisations, conform to the training, or at least feel 'powerless against it' (Hermes, 1972 in Elmes & Costello, 1992, p.428). Elmes and Costello describe communication skills training as a 'social drama', drawing attention to the actors involved and the staging of programme elements. Rituals, testimonials, and presentation of material all contribute to a certain dramatisation of the messages.

The above-mentioned studies contribute knowledge about the encouragement of confession and conformity, surveillance activities, and processes of subjectification, all of which contribute to participants' submission to certain leadership identities. Developing strategies of resistance is either stressed as important (Wilson et al., 2020), or considered nearly impossible to achieve, especially for student participants (Ferry & Guthey, 2021). Some critical scholars discuss opportunities of agency and resistance in leadership development programmes in general (e.g. Nicholson & Carroll, 2013), but few have empirically explored participants' reactions and counter-strategies to the process of being measured.

Contributing to our knowledge of quantification and its effects, a body of literature, primarily within interpretive accounting research, is concerned with the performativity of numbers in an organisational context. For example, Fauré, Cooren and Matte (2019) examine what numbers require in order to 'speak'. The authors argue that numbers' capacity to do things, to perform, is tied to their interaction with those using and responding to numbers. For instance, the authors point to the authority numbers assign to people and that people assign to numbers, and the work done to make a number matter in the first place. As the authors conclude: 'If nobody listens, cares or believes what the numbers are supposed to say, numbers will remain lifeless figures' (p.354).

Similarly, Dorn (2019) questions measurements' ability to more or less automatically prompt reactions, arguing that the reactivity of numbers depends on how social actors make sense of them. Exemplified with hospital rankings, Dorn refers to the variety of ways hospitals respond to their ranking, and makes the case that a measure, such as a ranking, does not necessarily put 'irrefutable pressure on organisations and [work as] an influential device in their evaluative environment' (p.343). These studies draw attention to the multiplicity of ways

numbers can be contextualised, used, and responded to, and how these elements determine if and in what ways numbers are performative.

Related to the performativity of numbers, Espeland and Stevens' (2008) conceptual paper has served as one of the main inspirations for my study's positioning and development. The authors appreciate here both that quantification has performative effects *and* relies on context. Drawing on the work of philosopher John Langshaw Austin (1962; 1975), Espeland and Stevens make the case that numbers perform different acts by guiding our attention and ultimately by making us do things. Since I likewise draw on Austin's work in my discussion chapter, a short account of his notion of the performative will be useful, before I then return to Espeland and Stevens' paper and outline how my study complements and extends the authors' perspectives on the performativity of quantitative assessment tools.

In his book, *How to do things with words* (1962), Austin's central argument is that by uttering words, we do not simply describe, we actually do. Austin distinguishes three kinds of speech acts: locutionary, illocutionary and perlocutionary. Of specific interest in the context of this thesis are illocutionary and perlocutionary acts.

Illocutionary acts are sentences that realise their intent through their mere utterance: sentences whose meaning is achieved through the act of saying them. For instance, uttering the words 'I hereby pronounce you man and wife' in a wedding ceremony, actually unites two people into marriage. The meaning of the sentence is made true in the moment the sentence is spoken. Examples of illocutionary speech acts also include the use of performative verbs such as to promise, to order or to request. For instance, when someone says 'I promise I will return the money to you', s/he has in fact made a promise.

Perlocutionary acts, in contrast, refer to the consequences or effects of an utterance. For example, if someone inside a room says: 'It's really cold in here', this could imply a request for someone to close the window thus performing an illocutionary speech act; making a request. What this utterance then initiates, for instance that someone closes the window or argues that it is not cold at all are the perlocutionary effects of the implicit request.

Perlocutionary acts thus refer to what is set in motion by illocutionary speech acts. Related to perlocutionary effects, Austin emphasises the numerous potential outcomes of an utterance, which can be both intended and unintended. In other words, the listener can react to the speaker's speech act in various ways.

Importantly, illocutionary speech acts are not always successful, meaning that they do not always realise themselves. In order for speech acts to be complete, successful or 'felicitous' (happy) as Austin terms them – successfully realising what they intend – a set of circumstances must be in place. If such conditions are not in place, speech acts risk 'misfiring', meaning that what is intended by the illocutionary act will not be realised (Austin, 1962, p.14).

Of these conditions, Austin considers authority relations to be particularly important. The speaker must have the authority to achieve what is intended by the spoken words. Declaring two people married requires the authority for such an act. In other situations, the authority is less official and ceremonial, but the receivers of the speech act must still recognise it. Successfully declaring a meeting adjourned, the person declaring must have the authority to do so (formally or informally). This means that the other people present must recognise this person as someone who can, and perhaps is expected to, make such a declaration. Also influencing the felicity of illocutionary acts are the physical surroundings. One must stand in front of two people getting married, when marrying them, attending a form of meeting that can be adjourned.

If we now return to Espeland and Stevens' paper, we may ask what it means for numbers to be performative. Thinking of quantification as 'speech acts' means that we look at how and in what ways numbers perform or do things. Espeland and Stevens (2008) argue that we turn our attention to quantitative tools' persuasive, reactive, and productive abilities. According to the authors, quantification is performative by 'intervening' in our world, creating or reinforcing social categories, enabling judgement, evaluation, and overall discipline. More specifically, they suggest, for instance, that we look at the authority we ascribe to numbers, enabling their persuasive force, and the interpretation and decoding efforts quantification and numerical pictures (graphs and models) call for. As a consequence, considering quantification's perlocutionary dimensions, Espeland and Stevens (2008) urge studies to be, 'sensitive to context' (p.404), so as to explore the purposes and meanings of quantification, and what quantitative tools set in motion.

Even though the authors point in the direction of more contextual perspectives on the effects of quantification, and emphasise how numbers rely on the authority we grant them, the authors remain primarily within the borders of quantifications' material requirements. In terms of the 'work' which quantitative activities demand, Espeland and Stevens (2008) refer to the skills, time, money, and coordination efforts which grading essays requires. Skills, time, money, and

coordination efforts are quite concrete necessities for quantitative activities to be carried out and implemented in the first place. The elements tell us something about what quantitative activities need in order to actually take place, not what the activities need in order to have an effect on the world. Moreover, Espeland and Stevens' acknowledgement of the context of quantification mainly involve the outcomes of quantification. The authors allocate most of their attention to the effects of quantification, its reactivity, ability to '[intervene] in the social world it depicts', and 'cause people to think and act differently' (p.412), and thus contribute with knowledge of the ways quantification does things or make individuals do things.

Likewise borrowing from Austin, my study complements Espeland and Stevens' (2008) insights, by taking their argument even further and putting the social context of quantitative activities to the centre of study and discussion. In terms of the 'work' quantification requires, I will argue that numbers rely on certain conditions (beyond skills, time, and money) in order for them to perform their intentions, which is why I specifically look into how test practitioners work to create and maintain these circumstances, and how test takers respond to this. My study thus extends Espeland and Stevens' (2008) work with empirically informed insights, particularly about the role of authority granted to quantitative measures and their advocates, and the mediating work quantitative tools require to have performative effects.

In sum, my study represents a critical, contextual take on testing, by exploring the research questions: *How do leadership measures rely on and promote norms? And how do social actors in organisations reinforce or undermine these norms?*

METHODOLOGY

In this chapter I present my methodological choices and reflections and cover how I have approached the field of leadership measures in a way that deepens our understanding of the norms within and around such assessment tools. The chapter includes reflections on the ontological and epistemological traditions and perspectives my study rests on, since these steer my subsequent choices regarding research strategy, concrete methods, and theorisation approach.

Research perspective

According to Denzin and Lincoln (2000), 'every researcher speaks from within a distinct interpretive community that configures, in its special way, the multicultural, gendered components in the research act' (p.18). Accordingly, I approach the research field with a particular set of ideas and a distinct point of view, affecting my perception of the studied phenomena. In the following, I explain with what particular mindset I have approached the field of leadership measures.

My research perspective lends itself to the constructivist (Berger & Luckmann, 1966), interpretivist tradition. What is of interest is how concepts such as leadership and norms are constructed and maintained through practices and language. Consequently, I focus on subjective experiences and meanings attached to objects such as leadership measures. Since, as Morgan (1980) expresses it, 'what passes as social reality does not exist in any concrete sense, but is the product of the subjective and inter-subjective experience of individuals' (p.608), the aim is to uncover how social actors make sense of, construct, and negotiate the four measurement tools and the activities surrounding and supporting the instruments.

This position provides an opportunity to understand the use of tests in leadership development as a value-laden practice and offers a counterpart to the reductive functionalist and quantitatively based studies that dominate the field. Since,

'understanding meaning and intentionality is emphasised over and above causal explanations' (Prasad, 2005, p.14), my interest lies in how individuals make sense of and influence particular leadership measures.

Accordingly, I view test effects as depending upon individual interpretations and reactions, therefore accounting for contextual influences and complexities. Using this framework, subjective nuances, ambiguities, and paradoxes are brought out, offering an alternative understanding of contemporary use of leadership measures compared to that of quantitative studies.

In other words, to counter the reductionism of such quantitative instruments that capture the observable and risk neglecting the constructive qualities of the world, my approach is informed by critical, interpretivist perspectives; embracing complexities, nuances, and multiplicity. More specifically, the aim of this study is to generate 'insightful descriptions' that 'present the phenomenon in new and revealing ways' (Hammersley 1992 cited in Bate, 1997, p.1168). Instead of taking leadership measures at face value, I consider the tools as expressing something other than an objective truth and working in other ways than perhaps initially (or officially) intended.

Quantifying human behaviour, competencies, and reputations represents certain convictions, beliefs, and discourses, which calls for critical questioning. The demand for and implementation of tests are historically and socially conditioned and created. Consequently, a test contains and promotes much more than what it purports: a set of items measuring the level (of effectiveness) of certain competencies. From a critical, interpretive position, tests are based on normative beliefs about leadership, formed by historical and social events, research streams, and commercial trends. The theoretical analysis of the tools themselves therefore intends to reveal the, at times hidden, normative assumptions embedded in leadership measures. Following this perspective, leadership measures are interesting in terms of their ideological and political properties. Studying leadership measures critically also means avoiding the reproduction of dominant ideologies or taking the legitimacy or naturality of such measures for granted (Alvesson & Sköldberg, 2009), countering what some organisations tend to do, namely treat tests as a natural, perhaps even unavoidable, choice, with quite unproblematic and predictable implications.

Leadership measures convey the idea that leadership is a consistent phenomenon which can be categorised, that patterns and units exist, and that language in a test mirrors social reality. In contrast, I argue that such measures and their language represent assumptions, values, and belief systems, and that their effects are both

unpredictable and influenced by numerous contextual factors. Most importantly, the act of commensuration, reducing leadership and personality to numbers, can be seen as a form of silencing (voices, nuances, inconsistencies, experiences), that a critical, interpretative approach counter by voicing and foregrounding interpretations, paradoxes, and potentially problematic assumptions.

Studying an evaluative quantitative method such as leadership measures or personality tests in a critical manner calls for different, contrasting methods than those that have been used to construct the tools. In other words, since the method I explore is quantitatively based, the methods I employ in this study need to offer something else than the studied method itself claims to offer. Test takers' experiences of being tested cannot be captured by using the same method (e.g. quantitative tests or surveys). Also, quantitative studies on test experiences and effects typically restrict their focus to that of test takers, without consideration for other social actors involved in the measurement process, such as feedback givers or those who permission the test. I argue that, first, test experiences and effects do not lend themselves well to being expressed numerically or placed in a pre-defined box, and second, that the effects of tests can be best understood if we include a more contextual perspective.

In my study, this approach means that my observations, interviews, and document analysis represent specific kinds of realities, whose uniqueness is valuable, adding to the nuance and complexity of the studied phenomenon. The empirical analysis of social actors' test understandings and experiences of tests thus reveals a richness and depth that is more limited in quantitative or positivist leadership studies. For example, I view the individual test taker, not as a project for improvement, or a simple test recipient, but as a person with particular and unique opinions about the test, affecting the subsequent lives of the measures.

While acknowledging and highlighting the variability and diversity of both the test takers' experiences and the four measures, I also bring out patterns and regularities. These speak to the tendencies within test use, suggesting that even though the measures are different, they, in some respects, represent the same phenomenon.

Case study

Setting out to generate an in-depth and multi-faceted understanding of test use in leadership development, the study is built on cases or occasions of observation, as I term them. Exploring test use this way means that the particular and unique is foregrounded (Stake, 2000). My study is a mix between an 'intrinsic' and 'instrumental' case study (Stake, 2000). It is 'intrinsic', as I focus on the case itself, i.e. the particular leadership measure. The specific tests, their particular qualities, formulations, and formats are of interest in themselves. At the same time, the case study is 'instrumental', in that I explore the particular measures in order to gain insight into an issue and to better understand a phenomenon: the use of quantitative assessment tools (in leadership development). The studied tools serve as examples of leadership measures, and could therefore have been different ones.

Since the purpose of my study is to generate insights into tests' design and how they are framed and experienced, with a focus on norms and normativity, I have studied four measures. This allows me to identify similarities, differences, tendencies, and patterns, while still foregrounding each measure's uniqueness. The intention is not to formally generalise, but to offer new insights through 'the force of example[s]' (Flyvbjerg, 2006, p.228). However, the similarities of these measures and how and why they are implemented speak to tendencies possibly beyond the context of leadership measurement.

Four measures – and their empirical contribution

The four quantitative assessment tools are named: 'The Extraordinary Leader', 'HD Leadership', 'Hogan Leadership Forecast', and 'People Test Person'. These are either specifically targeted at managers ('The Extraordinary Leader', HD Leadership', and 'Hogan Leadership Forecast'), or possible to use in both recruitment and development purposes ('People Test Person').

Although I acknowledge how companies' structure, management, and culture influence how assessment tools are experienced by test takers, the core empirical material for this research comprise the occasions of observation and the assessment tools themselves. I therefore provide here descriptions of the specific measures and the tools' contribution to my study.

The Extraordinary Leader

This is an American, commercial measure developed by Jack Zenger and Joseph Folkman. The tool is based on a 360° feedback model, consisting of feedback from peers, employees, and supervisor(s), in addition to a self-evaluation. In the setting where I studied the tool, the respondents consisted of four peers, six employees, one manager, and one test taker. Each respondent, including the test taker in question, rated the person on 49 items that together purport to measure 16 competencies, on a scale from one to five.

The Extraordinary Leader is meant to provide test takers a picture of their overall 'leadership effectiveness' (PowerPoint slide from the workshop). The report also shows if test takers have any 'fatal flaws' by which is meant strong negative feedback indicating 'below average capability in an area that is mission-critical to [their] job' (PowerPoint slide from the workshop). Fatal flaws are expected to lead to 'performance problems, career plateaus, job failure, and damaged relationships' (PowerPoint slide from the workshop).

The rationale behind the tool is that 'peak performance can be engineered' (The Extraordinary Leader - Participant Manual, 2015, p.6, module 6). This mechanistic logic leads to the recommendation that so-called leaders develop three to five competencies with a rating for each which lies within the 90th percentile in comparison with the rankings for other leaders. As Zenger Folkman states: 'You don't have to be perfect' (PowerPoint slide from the workshop), you just need Profound Strengths in three to five areas (Zenger & Folkman, 2017). Reflecting a concern with causality, the conviction is that by developing these profound strengths the organisation and leader will 'truly flourish' (Zenger & Folkman, 2012, p.xii). The research supporting this logic is based on data from around 20,000 leaders, who have been measured with different 360° feedback instruments. Collectively, the data were comprised of 1,850 survey items describing different behaviours. The test developers' subsequent analysis revealed 16 competencies that differentiated 'the best from the worst' (Zenger & Folkman, 2012, p.5). Moreover, from the data, the developers derived 49 items that 'accurately measure leaders' effectiveness at these specific competencies' (Zenger & Folkman, 2012, p.5). The 16 competencies are organised in five behaviours that comprise the 'Leadership Tent Model'. For an overview of the instrument's items, competencies, and behaviours, see Appendix 1.

I studied this tool at PharmExtra (pseudonym), a health care company which started using The Extraordinary Leader in 2016. According to the company's Leadership Specialist, Aaron, the company deliberately chose a 360° leadership

assessment tool, since they consider tests based solely on self-reporting to be inadequate. The belief is that with a 360° model, the test taker will get information about how they are perceived by different people, providing a fuller and more trustworthy picture.

The tool is used as part of a virtual workshop within a leadership development programme for second level managers (managers managing managers). The measure is introduced to the participants by an external consultant. At the time I did my observations at PharmExtra, the company had held this workshop four times, with no equivalent tool prior to this.

Studying The Extraordinary Leader has provided me with insights into the introduction and framing of leadership measures. I attended a workshop where the tool was introduced for 2.5 hours. This was an opportunity to hear what and how certain points, such as the tool's validity and trustworthiness, were emphasised by the consultant, and how this was received, within the restrictions of the virtual workshop, by the participants.

HD Leadership

HD Leadership is a commercial personality test developed by four Danish psychologists. It consists of 271 items. The company, Human Developers, was founded in 2013 and its stated aim is to 'provide an all-in-one solution for HR in the business sector, in terms of both assessment and development' (Human Developers, 2023, n.p.).

On Human Developers' website, their 'test system' is described as being:

> …based on our effective business psychology personality test, which measures personality accurately and nuancedly. It gives an in-depth picture of the person's characteristics, dynamics and potentials, and reflects the personality profile in relation to the current situation of the test subjects, including the purpose of the testing. (Human Developers, 2023, n.p.)

The test consists of 26 'personality scales' that are uncovered by the 271 items, which are answered on an ordinal scale with five possible answers: 'highly agree', 'agree', 'from time to time', 'disagree', and 'highly disagree'. Items include:

> When assessing an issue, you always need to first consider the concrete.
>
> Very rigid principles often strangle innovation.

According to the test developers' PR material, the scales should not be understood as measuring more or less permanent traits. On the contrary, the belief is that a person's 'properties, dynamics, and potentials' are changeable, especially if the person undergoes coaching sessions. The 26 scales in HD Leadership are arranged into four categories: 'Personal Strength' 'around [which] are the three roles: 'The Controller, The Inspirator, The Strategist' (HD Leadership, own report). The generated report goes through all 26 scales, sorted by their deemed relevance for one's personal strength, role as controller, inspirator or strategist.

I studied HD Leadership primarily from the test takers' angle with access to a Danish company, Logistica (pseudonym) who uses this personality test mainly for managers or potential managers who are participating in a talent programme. I interviewed 11 test takers, one of the test developers, and a test administrator. Also, I took the test myself.

Studying HD Leadership has contributed insights into aspects of test development but also the more practical side of test use: how a tool is used to both develop existing leaders and assess the 'leadership potential' of employees participating in a talent program.

Hogan Leadership Forecast

This self-assessment tool, developed by Joyce and Robert Hogan, consists of three parts: 'Hogan Personality Inventory', 'Hogan Development Survey', and 'Motives, Values, Preferences Inventory', generating three reports: 'Potentials', 'Challenges', and 'Values'. By answering 150-200 questions per assessment, the tool is aimed to inform how others perceive you, ultimately providing what is described as 'strategic awareness':

> It's vital that people understand the difference between the way they see themselves and the way they are seen by their peers, managers, and direct reports. As the industry-leading expert, we've developed solutions that provide critical insight into characteristics that not only facilitate an individual's success, but, more importantly, can cause failure and career derailment. (Hogan Assessments, online, n.p.)

The Hogan Leadership Forecast is not a multi-rater assessment. However, the report is designed to tell test takers how others perceive them, solely based on self-assessments.

The ability to infer how others perceive you based on your self-report is made possible due to studies where employees have been asked which of a certain

number of provided characteristics describe their manager the best. These results are then considered against the manager's test results. By pooling thousands of these together, a person's Hogan profile is connected to characteristics that other people, according to the studies, will likely ascribe to this kind of profile.

According to Hogan's official website, the Hogan Leadership Forecast:

> …provides an in-depth look at a leader's performance capabilities, challenges, and core drivers. Use it for succession planning decisions and leadership development, and you'll see your current and future leaders excel. They'll gain strategic self-awareness to leverage their strengths, avoid behaviors that get in the way of success, and gain insight into the culture they create for their teams based on their motivators and values. (Hogan Assessments, online, n.p.)

Two parts of the Hogan Leadership Forecast are based on a five-point Likert scale: 'strongly disagree', 'disagree', 'agree' and 'strongly agree'. The Motives, Values, Preferences Inventory, comprises items with three options: 'agree', 'disagree', or 'don't know'. Items include for example:

> I am a very confident person.
>
> Fear has been a big driving force in my work.
>
> I am proud to be someone who follows others.

The Hogan Leadership Forecast generates three main reports: 'Values', 'Potentials', and 'Challenges', a summary and a 'Flash Report', which gives an overview of the results of all three main reports (see Appendix 2). All the reports start with definitions and cues on how to use the information.

I was able to observe the use of the Hogan Leadership Forecast at BigBank (pseudonym), where I conducted interviews and observations at what the company called a 'community meeting'. At this meeting, the rationale behind using Hogan and the choice of eight specific competencies as the most important ones, were presented by the company's Transformation Consultant.

I interviewed two external consultants, Jacob and Miles, the Vice President, James, responsible for the leadership development programme which the Hogan Leadership Forecast was a part of, and the Programme Director, Megan, who had chosen to purchase and implement the tool at BigBank. Following the organisation's request that I did not approach any test takers myself, my options were rather restricted, and only two test takers contacted me. To partially

compensate for the limited insight this afforded, I took the test myself and received feedback from one of the consultants.

People Test Person

People Test Person is a commercial and normative self-assessment tool developed over the past 15 years by the Danish company People Test Systems. On People Test Systems' website, they state:

> People Test Person is a well-documented personality test which is applicable for, among others, recruitment, personal development and value-creation for employees and leaders. People Test Person provides a thorough and nuanced image of the candidate's personality and behaviour, by covering 12 overall character traits and 36 subjacent qualities. The test also measures to what extent the candidate takes responsibility, has a constructive attitude and a realistic evaluation of their own abilities. (People Test Systems, online, n.p.)

After taking the test, an analysis is generated where the test taker's results are compiled to form four categories of personal attributes: personal characteristics, dynamic characteristics, qualitative characteristics and cooperation. For each category, three character traits are listed (making 12 in total), which are then further divided into three qualities (36 in total).

I studied this tool in terms of how it was introduced to practitioners. I attended a certification programme, where people from different companies learned how the tool was built, its purpose, how to give feedback, and what ethical guidelines to consider. I also interviewed two people from People Test Systems. Emma, who is involved in developing the test and assuring its quality, and Elizabeth, who is primarily responsible for certification programmes, assessments, and advising clients. I did not interview any test takers, since my access was to the company developing the tool, and not an organisation using it.

The empirical contribution of these observations was insight into test certification; how test practitioners teach and communicate test theory, and how participants receive and work with this information. This access offered insight into test framing at certification workshops where future test practitioners are trained or moulded, so to speak, to represent and use the tool in a certain way.

Differences and similarities

The four tools differ in various ways. The Extraordinary Leader is a multi-rater tool, whereas the others are based on self-assessment. The Extraordinary Leader

and Hogan Leadership Forecast were developed in the United States and focus on behaviour and other people's perceptions of the test taker. HD Leadership and People Test Person originated in Denmark and concentrate on personality and self-perception.

Besides measuring different things, the tone, construction, and formats of the tools vary. The Hogan Leadership Forecast for example includes these items:

> I know why the stars twinkle.
>
> My friends know how to party.
>
> As a kid I often wanted to run away from home.
>
> I have felt bitterness towards my parents.

In contrast to these quite evocative and emotional items, the items in HD Leadership have a different, more formal tone:

> When you are thinking about the details, you consider the different parts of an issue.
>
> Practical sense is more important to judgements than abstractions.

No items on family ties or emotional matters are included in HD Leadership.

The Extraordinary Leader contains a lot of implied meanings and taken for granted expressions such as:

> Works hard to "walk the talk" and avoids saying one thing and doing another.
>
> Balances "getting results" with a concern for others' needs.
>
> Has a perspective beyond the "day-to-day" work to take a longer-term, broader view of business decisions.
>
> Frequently encourages others to consider new approaches and ideas (e.g., avoids getting stuck in a "one right way" approach).

According to the test developers, these items are observable and possible to evaluate from the outside. The self-assessment tests, on the other hand, require personal, more intimate, answers.

Despite some obvious differences, the four tools are similar in quite a few ways, suggesting that the test developers share some ontological and epistemological

assumptions. For this reason, I was able to also compare the measures, despite their differences.

First, the measures are all developed with the purpose of measuring some 'quality' such as personality, leadership effectiveness, behaviour, or a person's image. Second, the reliance of fixed choice formats suggests that test developers believe that they have ensured conceptual equivalence, that is, that respondents understand the meaning of questions and expressions identically. Essentially, this means that individual interpretations, states of mind, and previous test experiences are variables that will not affect the test results significantly. Third, another shared assumption is that a person's leadership is improved through self-awareness, either by accessing other people's perception of you, or by having your self-assessment analysed, structured, and translated through a leadership measure and feedback giver. This rests on the conviction that the more aware one is of one's strengths, weaknesses, and what impression one leaves with others, the better development foundation one has. The self-assessment tests presuppose self-aware test takers in order for them to respond to test items in an honest way from the beginning. The tests thus assume a certain level of self-awareness, while claiming that it is also the product of the test.

Besides the above-mentioned assumptions, the four measures are also intertextually connected. Albeit officially subscribing to different psychological theories, and using different terminologies, the measures target a number of similar or even identical characteristics (norms), suggesting that they draw on some of the same psychological theories and leadership philosophies. For example, a test taker's 'sociability' or so-called 'extraversion' is included in all four measures. Other characteristics that occur in two or three of the measures are 'focus on details', 'empathy', 'being innovative or visionary', 'dominance', 'risk taking', 'self-confidence', and 'responsibility'. These are the most direct or apparent overlaps.

There are other, less obvious, overlaps. The developers use different terminology, making their similarity to other measures harder to identify. For instance, in Hogan Leadership Forecast, 'adjustment' is a category that concerns 'composure, optimism, and stable moods' (Hogan Potentials report), whereas People Test Person measures 'stability of mood' and 'optimism' separately.

The fact that all four tools include a person's level of extraversion, which is a Jungian term, suggests that the measures are intertextually connected, in that they indirectly refer to some of the same psychological theories that share many of the same assumptions. Personality tests thus target some of the same characteristics

and test developers' choices are, as most academic or scientific work, based on pre-existing knowledge, psychological theories and previous tests.

Summary

The variety within the four studied tools offers valuable insights into the different leadership assessment tests on the market. Representing different belief systems, the instruments draw attention to the diversity in and range of leadership measures. However, in spite of the tools' differences, there are also several similarities and alike mechanisms.

Given the diversity represented in these tools, the various types of access I had to them, and the nature of my empirical material, the tools contribute different insights and are therefore not equally represented in all sections which follow in later chapters.

The different types of access and empirical material; attending workshops and a certification programme, taking tests myself, and interviewing different actors with different interests, mean that I have gained insight into test use from various angles. I have covered different perspectives of the test process: that of test takers, of consultants promoting and facilitating the test (and feedback), and of test developers.

Approaching the phenomenon from these angles has allowed me to identify paradoxes, ambiguities, and complexities that might otherwise be concealed. More specifically, by giving voice to different actors about the same phenomenon, it has been possible to identify and explore similar and contrasting experiences and beliefs. As a result, tendencies, patterns, and paradoxes in test practice come forth.

A multi-method approach

In order to generate rich descriptions of how leadership measures are framed and experienced, I have used multiple methods. Observations, document studies, and interviews allowed me to approach and understand the phenomenon from different angles (Stake, 2000). Following Denzin and Lincoln's (2000) accounts of qualitative research, a variety of methods contribute different perspectives and insights, supporting and contrasting each other in the attempt to reach 'rigor, breadth, complexity, richness, and depth' (p.5). In the case of leadership measures,

where several interests are at play, these, sometimes diverging, perspectives have contributed a nuanced understanding of the phenomenon.

My choice of using multiple methods is also based on the epistemological implications of my understanding of leadership measures. As covered in this chapter in the section called 'Research perspective', I have approached the phenomenon from an overall constructivist and interpretivist position, implying that my aim has not been to uncover some essential truth about leadership measures. Instead, the interest lies in the ambiguities, inconsistencies, and paradoxes, which is why employing a variety of different methods has been helpful.

The individual reasons for each method are given in the next sections. For an overview of my empirical material, see Table 1.

Table 1: Overview of empirical material

Measure (organisation)	Texts	Observations	Interviews
The Extraordinary Leader (PharmExtra)	A 360° sample report The book, *How to be Exceptional* (2012), by Zenger and Folkman PowerPoint presentation from the virtual workshop Website material	Two virtual workshops and almost three weeks at the Training and Leadership Development department Coaching session on Skype between Michael (consultant) and Joseph (test taker)	17 with test takers One with Aaron who chose and administers the test One with Michael, the external consultant who facilitated the virtual workshops
HD Leadership (Logistica)	Own test reports Educational material (for certification purposes) Handbook Popular articles and PR material Website material	Own test feedback session with Cathy	11 with test takers One with Julie, facilitating the test in Logistica One with Cathy, an external psychologist, consultant, and test developer
Hogan Leadership Forecast (BigBank)	Own test report PowerPoint presentation from the community meeting Website material	Community meeting, where the role of the test was explained to future test takers Own test feedback session with Jacob	Two with test takers Two with external consultants currently working in different agencies One with Megan, Programme Director, responsible for test selection One with James, Vice President, responsible for the leadership development programme in which the test played a part
People Test Person (People Test Systems)	A sample test report Educational e-learning material (for certification purposes) Manuscript on how to give feedback Website material	Test certification workshop	One with Emma, responsible for documentation, assurance of quality, and test development One with Elizabeth, responsible for 'delivery', the educational programme, assessments, and consultancy

Texts

Texts, in the form of the tests themselves, test items (questions), test reports, website material, and educational material, have played a significant empirical role in my study. It is primarily through document analysis that I respond to my first research question: *How do leadership measures rely on and promote norms?*

I consider texts to be not only representations of ideals, principles, values, and norms, but textual agencies promoting norms and ideas of normalcy (Cooren, 2004). Despite test reports' autogenerated texts, with their buried or distant authorship, the tests speak on behalf of the company and test developers who created them. I have therefore interpreted the items of the tests as representations of larger dimensions with underlying normative significance and as examples of assumptions on which the test is built. Take for example, the test item: 'I am proud to be someone who follows others' (Hogan Leadership Forecast). In this item, strong assumptions are clearly present about the dichotomy between leaders and followers, and that pride is likely to be involved. Moreover, I have given focus to the distinctive use of styles and visual formats such as graphs and tables, as these elements can indicate the intention to normalise and regulate test takers' behaviour.

According to Atkinson and Coffey (1997) these documentary realities 'create their own hierarchy and legitimate authority' (p.58). Combined with rhetorical devices and the absence of clear authorship, the tests appear authoritative, official, and factual and imply that the reality constructed exists independently (Atkinson & Coffey, 1997). These techniques are important to bring to light when analysing test material, since a test is a document designed to appear convincing and valid. I have therefore read the tests as texts that construct their own realities, 'according to conventions that are themselves part of a documentary reality' (Atkinson & Coffey, 1997, p.61). This means analysing the test as a text, and as part of a wider system of texts, with an authorship and, perhaps implied, readership.

Besides the measures themselves, my material consists of website material (all four measures), educational material (People Test Person and HD Leadership), promotional articles (HD Leadership), a Handbook (HD Leadership), PowerPoint presentations (Hogan Leadership Forecast and The Extraordinary Leader) and the book *How to be Exceptional* (2012), which provides the basis for The Extraordinary Leader.

Observations

On the seven different occasions I made observations, where tests had a central role, I have seen how social actors present, frame, and respond to leadership measures. It is exactly in these types of social settings where actors play different roles and support, moderate, or obstruct the normalising potentials in the measures, that I got an understanding of *how social actors in organisations reinforce or undermine the norms within the measures.*

I understand the process of observing as a theory-laden undertaking (Hanson, 2000), shaped by prior knowledge of the observed and the language used to describe and explain what we know. Having previously worked at HR departments where personality tests laid the groundwork for recruitment, leadership and team development, I have gained insights into the role and influence tests can have. Moreover, I have acquired knowledge about historical developments of measurements, experimental psychology, and the emergence of concepts such as 'norms' and 'objectivity'. This knowledge enabled me to (or perhaps, at times, restricted me from) seeing and, more importantly, interpreting things in a certain way. In accordance with Hanson (2000), I therefore consider it my task to show how 'data are molded by different theories or interpretations or intellectual construction' (Hanson, 2000, p.175). This will be further unfolded in the section called 'Analytical strategy'.

I have termed my observations 'occasions' to focus attention on the events themselves which occurred in a limited time and space, rather than the specific sites where they took place.

Attending situations where tests played a key role gave me the opportunity to observe the measures outside constructed interview settings. These situations would have taken place regardless of my observation and attendance. I therefore gained insight into everyday test settings: how tests are talked about, how test theory is taught and communicated to practitioners and how it is referred to in informal conversations.

I made observations in connection to three out of four of the measures, excluding on site observations at Logistica. However, being tested myself also served as an opportunity for observation. Even though the test subject was myself, getting feedback on my test allowed me to observe a feedback session. I therefore paid attention to my own internal emotional process, what I was being told, and how I was being told this. For instance, I observed how the consultant put together a

narrative, emphasised or ignored certain points, and asked me particular questions.

At all the occasions of observation, I introduced myself but then refrained from actively participating. This was because I wanted to gain insight into the ways the tests were presented and received without my interference. Also, the workshops and meetings followed a planned and rather strict schedule, leaving little room for interruptions. At the workshop where The Extraordinary Leader was introduced, there was a formal process for participants' questions. Three options of communication were available during the workshops. Participants could enable their microphone (the majority of the time, only the consultant's microphone was on in order to eliminate any background noise), digitally raise their hand by clicking a button named 'raise hand' (see Figure 2), or write their questions or comments in a shared chat that everyone could see.

Figure 2: Activation buttons, virtual workshop

I also avoided interfering since the phenomenon under study is personal or even emotional to the subjects, potentially making interferences disruptive or upsetting. Lastly, I aimed for a certain amount of distance, since I did not want my previous professional and personal experiences with tests to influence my interpretations too significantly. That said, I have also considered my own experiences with quantitative assessment tools to be a potential advantage, in the sense that they have enabled me to better understand test practice and perhaps the test takers' individual experiences.

Similarly to my approach to observation at PharmExtra, described above, at People Test Systems and BigBank I attended the certification workshop and community meeting, respectively, as a somewhat passive participant.

The individual occasions of observations are presented below.

First occasion

The first observations were on April 30th and May 3rd 2018 at PharmExtra, where The Extraordinary Leader was introduced to test takers. The workshop was based on Zenger and Folkman's book, *How to be Exceptional* (2012). Each participant was provided with the book and a participant manual. Prior to the workshop, the participants filled out the 360° assessment themselves and asked the required number of other people to fill this out.

The workshop took place at the company's headquarters. On April 30th, the external consultant spent 2.5 hours introducing the measure. He did this twice; for one group of participants during the morning and one group in the afternoon. At the second workshop on May 3rd, the session began with a repetition of key points from the previous one. After this followed more specific exercises and discussion about the participants' results and how they were going to work with these. The external consultant was responsible for both sessions, and he also gave the test takers their test feedback in individual coaching sessions.

Participating at this virtual workshop gave me an opportunity to observe how a tool such as The Extraordinary Leader is framed by an external consultant, in other words, what the consultant emphasises and how. Being virtual, the workshop was recorded, so besides attending it in real-time, I also got access to the recorded sessions. Both the external consultant, Michael, and I were present at the company's headquarters for the two days of the workshop, allowing me to ask him questions about the tool and the workshop.

After the workshop I spent two weeks at PharmExtra's headquarters where the Leadership Specialist, Aaron, provided me with an office space in the Training and Leadership Development department. I used this opportunity to learn more about how employees at the company work with leadership development, what role the assessment tool plays, and how they generally talked about and referred to the workshop. These observations enabled a deeper understanding of the norms ascribed to the workshop and assessment tool. This also gave me the chance to speak more informally with Aaron.

The second set of observations took place over four days in October 2018. My observations were related to informal conversations and everyday life at the office. Since some time had passed between May and October, I also used this as an opportunity to clarify what had emerged after having interviewed several of the workshop participants.

While at the department, my role was more active and therefore more potentially influential than at the workshop. I wanted to understand the reasoning behind the organisation's test use, and the thoughts, beliefs, and concerns related to it. Therefore, a certain level of closeness was needed. I initiated conversations (whenever I deemed it natural) and asked questions.

Second occasion

The second occasion of observation was at BigBank on March 18th 2019. Here, I participated in what was referred to as a 'community meeting' for managers who were interested in hearing about BigBank's plan concerning Hogan Leadership Forecast and the rationale behind choosing the eight specific competencies on which the managers were encouraged to be measured.

Around 40 participants were shown a PowerPoint presentation with information about the Hogan Leadership Forecast, the eight competencies, their alignment with the bank's own leadership model, and different network activities. Almost half attended virtually. At the meeting, these future test takers had the chance to ask questions about the process, why the bank had chosen to implement the Hogan Leadership Forecast in the first place, and how the eight competencies had been decided as being the most important ones. It was therefore an especially interesting opportunity to observe how those who had designed the leadership development programme and chosen to use this form of leadership assessment answered questions and framed the process.

Third occasion

The third occasion of observation took place on September 18th 2019 at People Test Systems during a test certification workshop. Here, participants became certified to give feedback on People Test Person. During the certification, participants practised giving feedback to each other. The workshop ended with an exam where they would give another person feedback on the entire test.

Prior to the certification workshop, participants had to pass an online exam after reading through e-learning material consisting of eight modules about different characteristics the test measures ('personal', 'dynamic', and 'qualitative'), ethical

guidelines, and questioning techniques. Likewise, I gained access to the e-learning material and tried to pass the exam. In the traditional teaching parts of the workshop, where the instructor gave a background of the test and the different measured characteristics, I sat at the same table as the participants. It was only in the group exercises and one-to-one feedback exercises that anyone could notice my different role. In these, I either observed the feedback process, conversed informally with a People Test Systems staff member, or structured my field notes.

Interviews

Semi-structured interviews with test takers have informed my analysis of the different responses to both the tests' normalising potentials and test practitioners' mediating strategies. Together with informal conversations, these interviews have added nuances to my observations, particularly those with test practitioners (consultants, test developers and test facilitators responsible for their organisation's test choice and use).

In total, I interviewed 30 test takers, eight consultants, and four individuals from the respective companies who have played a role in either choosing the particular measure or implementing it internally in the organisation. All of my respondents have been anonymised. 11 of the interviews were in English and 31 in Danish. For an overview of respondents, see Table 2.

Table 2: Overview of respondents

Measure (organisation)	Respondents		
	Test takers	Consultants	Test selectors/facilitators
The Extraordinary Leader (PharmExtra)	Martha, Freddie, Allison, Thomas, Dan, Vera, Carl, Sebastian, Joseph, Rachel, Oliver, Frank, Richard, Sean, Eric, Tim, Vincent	Michael	Aaron
HD Leadership (Logistica)	Connor, Gabriel, Layla, Samuel, Harry, Noah, Nathan, Leo, Catherine, Erin, Daniel	Cathy (also test developer)	Julie
Hogan Leadership Forecast (BigBank)	Charlie, Ethan	Jacob, Miles	James, Megan
People Test Person (People Test Systems)		Elizabeth, Emma (also test developer)	
Other consultants representing various tools		Violet, Eva	

Through interviews with test takers, I got an understanding of how they put their test experience into words. By encouraging them to talk freely about their experience, they described, evaluated, criticised, or praised the test and the test practitioners' strategies. From these accounts I have been able to detect different types of responses to both the norms within the tests and the strategies employed by test practitioners. For example, statements about how someone ought to act or perform based on their test's results, I consider supportive of my identification of the tests' normalising aims. Through interviews, test takers expressed responses that either adopted, questioned, ignored, commended, or resisted the normalising potentials in the tests. Concretely, I have treated statements including words like 'ought', 'should', 'must', 'best', and 'worst' to be empirical sources of normativity and tied to the tests' performative potentials to normalise.

It is also in the interviews that I have identified test takers' linguistic responses to test practitioners' mediating strategies. Test takers either explicitly comment on the work test practitioners do in order to frame the measures a certain way or express their experience with the strategies they have encountered in more subtle ways.

These semi-structured interviews allowed for a balance between prepared and follow-up questions (Kvale, 1994), to dive deeper into interesting themes that arose. Certain questions were part of every interview so as to compare answers and interpretations, but I generally used techniques that allowed for adaptability, detours, and elaborations. My interest was within the particular, which is why individual follow-up questions were necessary to understand the matters under study as deeply as possible.

Some interviews were quite structured, others more dynamic and fluid. In my interview outline, I alternated between asking broad, open questions and more specific ones. The broad questions let the interviewees answer as they saw fit, interpreting the questions freely and emphasising what they found important (or believed I would find important). This revealed what the test takers found memorable, interesting, or difficult, and thus indicated what they took away from the experience. I asked questions such as:

Tell me about your experience with the tool.

What was your first reaction when you received the report?

How was it introduced to you?

What were your expectations prior to taking the test?

Did anything surprise you?

Why do you think [company name] has chosen to use this tool?

What role do you expect the report to have (or have had) in your everyday life?

Could you imagine a leadership development programme without quantitative assessment tools?

Most of the interviews were conducted via Skype for various reasons. First of all, a number of test takers were located in different countries. Second, many requested the option themselves since they considered it more efficient and flexible. Third, all four companies use Skype (or similar) regularly. The respondents were therefore used to communicating virtually.

As well as these virtual semi-structured interviews, I visited some sites for a number of interviews. I therefore had the opportunity to see the physical environment, converse informally, and join some interviewees for lunch.

The interviews with consultants focused on the specific tool they advocated, their general testing and feedback provision experiences, their opinions on other psychological assessment tools, and the test industry in general. In these interviews, the consultants argued for the validity and reliability of the tool they represented. In contrast to the interviews with test takers, these focused less on personal experiences with tests, and more on the consultants' professional role as feedback givers or sales people. Examples of questions with consultants included:

What role do leadership measures play in leadership development, according to you?

What can [name of the tool] contribute with?

How does it differ from other tools on the market?

Is there anything it cannot tell us?

What does the ideal test process look like?

What do you do in order to ensure this?

Do you see any problematic aspects of test use?

Interviewing individuals who are responsible for their organisation's leadership development and consequently, test use, gave a better understanding of the reasoning and legitimisation of companies' test purchases. The people responsible argue for the necessity of these tools and their hopes for their effects. I asked questions such as:

What made you choose this exact tool?

What considerations did you have?

Why is it important to measure?

What is your impression of the test takers' reception of the tool?

What are the consequences of a 'good' or 'bad' outcome?

Almost all interviews were transcribed in full. Three different transcribers were involved (one native English speaker and two native Danish speakers). One interview was lost due to technical issues. During and especially immediately after the interviews, I wrote down thoughts, reflections, and very early analytical points. This served as an overview and a guideline for later sorting and coding.

In line with a constructivist ontology, I approached the interview situation as a form of negotiated accomplishment (Fontana & Frey, 2000), co-produced by the interviewee and myself rather than an opportunity to neutrally or objectively collect knowledge from informants (Kvale, 1994; Stake, 2010). I acknowledge that I influenced the situation through my framing of questions, interpretations, and signals. As Schwandt (1997) puts it, the interview is a linguistic event where 'the meanings of questions and responses are contextually grounded and jointly constructed' (Schwandt 1997 cited in Fontana & Frey, 2000, p.663). This further implies that I have interpreted my interviews with regard to the context in which they were conducted.

I have also treated the interview material with reflexive caution, as encouraged by Alvesson (2011). Interview accounts can be influenced by cultural scripts, discourses, impression management, moral storytelling, identity work, or political interest. At times they represent something co-created in a unique situation between myself and the person I am interviewing (Alvesson, 2011). That said, I tried to find a balance between reflexivity and pragmatism. When interpreting the interview transcriptions, I continuously considered alternatives and attempted to challenge my assumptions. Specifically, I applied different analytical categories (scripts, moral storytelling, identity work) to the same account to assess how the meaning of the text changed. Analysing certain interview accounts like this has sensitised and nuanced my interpretations.

Getting tested myself

During my field work, I took three tests myself: Hogan Leadership Forecast and two of Human Developers' tests: HD Profile and HD Leadership. I only received feedback on my Hogan Leadership Forecast and HD Profile. Besides sensitising me towards my respondents' stories and emotions, it gave me the chance to experience and observe feedback sessions with a consultant. These two feedback sessions differed greatly in every way possible, drawing my attention to the significance of consultants' approach, attitude, and style.

Taking tests myself allowed me to access and experience the items that comprise the tools. Investing myself personally this way means that the distance between myself and the studied phenomenon has been reduced. In these two experiences I respond myself to both the tests' normalising potentials and test practitioners' mediating strategies. I have felt the frustration with item formulations, feelings of uncertainty when choosing a response option, concerns about my test results,

expectations about the feedback session, the thrill of discovering things about myself, surprises, and disappointments. In the process of taking these tests; receiving the links to the tests, filling them out, waiting for feedback, and receiving feedback, I have written down my thoughts and impressions.

I have chosen to share my test results on different occasions throughout this thesis, exposing and displaying the tests' assumptions about my personality. I have made this choice hoping to make testing less conjectural and more relatable to the reader.

Reflections from the field

Approaching this particular field from a critical position has brought different challenges. Meeting and interviewing individuals in favour of or even representing and selling certain measures, arguing for their value and necessity, have both had interesting and challenging implications. At lunches, meetings, and in informal conversations and interviews, test practitioners would speak with such enthusiasm about the tools and their contribution, that any criticism, or even curious questioning at times seemed too provocative or confrontational.

Consequently, I have had to find a balance between challenging these taken for granted convictions while trying to avoid offending anyone (too much). This was also something I considered when doing interviews, in terms of how much I should reveal about the aim of my study. If I was interviewing a consultant, challenging their assumptions sometimes led to very honest and open conversations, where the consultant expressed doubts, concerns, and paradoxes about their own work. Other times, critical questioning was met with counterarguments or the claim that academics were simply naïve. These challenges have been a constant source of both frustration and motivation. The fact that my endeavour has been able to provoke some of the people I have encountered, shows how established and legitimate the measurement field has become. The value of contemporary test practice is so recognised that critical or even just curious questioning can be met with massive astonishment and defensive argumentation.

When I asked critical questions about the limitations of leadership measures, or the assumptions inherent in them, most practitioners would react with long periods of silence, vague answers, refer to complex calculations and studies, or

emphasise the inescapable, self-evident necessity of such tools. Being met with this kind of response confirmed my argument for conducting this study. Test practice has become a taken for granted, naturalised part of organisations. As my empirical material will reveal, questioning its validity, by test takers or myself, is usually either rationalised or ignored. There are, of course, exceptions to this reaction. A number of test takers and a few consultants expressed reservations and concerns about test use. Even so, these consultants tended to then explain how the tool they represent avoids these pitfalls, or how their handling of test results is ethical and considerate.

Analytical strategy

In this section, I describe the type of dialogue that I have had with my empirical material; how I have gone back and forth between my empirical material and my preconceptions, assumptions, and analytical categories. My analytical strategy thus covers the part of my study that concerns how I have sorted my material, made selections, reductions, identified patterns and tendencies.

Sorting and coding my material were influenced both by what the respondents themselves emphasised and by constructed analytical codes. I have therefore remained close to the empirical material, but also allowed more analytical codes to guide the sorting process (Charmaz, 2006). More concretely, I labelled different segments of data that either simply described what the segment was about (by using the respondents' own words, for example) or that was influenced by preconceived notions which took the form of more abstract categories (Charmaz, 2006). I identified these themes by reading through my material, looking for repetitions, similarities, and differences between statements (Ryan & Bernard, 2003). For example, when I noticed that many respondents described the tool as helpful and tangible, I started looking for statements that expressed the contrary. Some of the initial codes were: 'scepticism', 'the tool as helpful', 'described changed behaviour', and 'attraction of numbers'.

My constructed analytical codes were partially informed by the existing critical literature about measurements and norms and thus reveal some of my own assumptions about the implications of quantitative assessment tools. I have used these concepts or notions as points of departure to inform the interview questions and reflect analytically on my empirical material (Charmaz, 2006). This meant that I looked for taken for granted ideas and beliefs and implicit assumptions since

I expected these to be linked to normative perspectives. As mentioned, I took notice of normative words such as 'ought', 'should', 'must', 'best', and 'worst'. Also, I paid extra attention to statements that more indirectly revealed normalising effects at play, such as the description of changed behaviour or submission to what was perceived as 'normal', 'appropriate', or 'encouraged'.

During this, I started connecting the themes, selecting the most significant ones (Charmaz, 2006). I compared experiences, actions, and interpretations and found similarities, contrasts, and tensions to explain or understand the patterning of the first order coding (Maanen, 1979). For example, I noticed how almost all respondents emphasised the importance of numbers. Later, connecting this tendency to other themes, the *different roles* of numbers emerged as an interesting theme. I tried out different interpretations, presented them to colleagues or peers, and continually checked back with my data to see which interpretations were most substantial and empirically supported. This process was also influenced by a search for what was 'interesting' (Davis, 1971) e.g. opportunities for dominant assumptions to be contested, more specifically when 'what seems… is in reality…' (Davis, 1971)[3]. This involved challenging the prevalence of themes by looking for the converse or different, and discovering new or unexpected relationships. In order to construct and experiment with these 'interesting' relationships (theorisations), I found it helpful to allow the process to be fluid, intuitive, and even messy (Gioia, Corley & Hamilton, 2013). Inspired by Swedberg (2012), the process can be described as a 'playful', 'imaginative' and 'explorative' discovery phase where the purpose was to find what was significant and interesting, meaning that many ideas and interpretations attracted attention and that justifying them became a secondary exercise.

Interpretation and presentation

Studying the use of leadership measures through a constructivist lens, interpretation and reflection have been my main instruments. With this position follows the notion that subjectivity is not something to be avoided, rather it is considered a strength and a necessity to understand human activity (Stake, 2010). I relied on inferring and interpreting meaning, based on what I heard, read, and

[3] 'Reality' is used here as representing new findings that contrast or contest previous understandings. The term is not meant to indicate that the new findings are 'real' in an objective, stable, or universal sense.

observed. I, with my particular academic background, personal experiences, and set of assumptions, have been the instrument through which I could paint a picture of the test takers' reality (Denzin & Lincoln, 2000). The interpretive process then involves making sense of the empirical material and seeing something as something (else). As Bate (1997) argues:

> The story will never be a telling but a retelling, never a transcription but a translation. There really is no such thing as "insider out", only an ambition to get closer to the natives, and a commitment to learning something about their world and what they make of it all. (Bate, 1997, p.1160)

While emphasising the strength of subjectivity, I have sought to communicate my knowledge claims transparently. Through multiple methods I have strived to base my interpretations on a comprehensive, full, and detailed picture of the happenings in the cases. With that said, the meaning of the material depends as well on the reader's interpretation. Since the end result will be a painted picture, a told story, as it looks from my perspective, it is open to interpretation by others. This is not however a disadvantage:

> With an explication of the perspectives adopted towards an interview text and a specification of the researcher's questions to an interview passage, several interpretations of the same text will not be a weakness but a richness and a strength of interview research. (Kvale, 1994, p.157)

Based on this, I explicate my perspectives, assumptions, and questions, but then let the text be, allowing for multiple possible meanings and interpretations.

To sum up, the reflexive framework employed works as a way of continuously questioning and challenging the dominant assumptions, my own assumptions, the empirical material, and especially my interpretations of this material. I have sought to interpret my interpretations, so to speak, to embrace the ambivalence and complexity in my empirical material rather than succumbing to the temptation of presenting a coherent, credible, and convincing story supported by well-chosen quotations and observation points.

Summary of methodology

Setting out to qualitatively explore the norms in and around four measures from an interpretive, critical position, the study's main instrument is interpretation. I consider the tests themselves and the meaning attached to them as socially constructed, continuously reproduced, and negotiated. As a result, I have focused on interpretations, argumentations, and legitimisations of test use. By this means, the phenomenon of interest is turned into a 'series of representations' (Denzin & Lincoln, 2000, p.3), studied in its natural setting, in order to 'make sense of, or to interpret, [the phenomenon] in terms of the meanings people bring to [it]' (Denzin & Lincoln, 2000, p.3). Through observations, semi-structured interviews, and informal conversations, I explored how different actors make sense of and influence test practice. The process of collecting empirical material was characterised by my interpretive stance. The material is viewed as co-produced, since I inevitably influenced the field by my presence, specific formulation of questions, and assumptions. The presented interpretations, the story of the story, is therefore a result of looking at the phenomenon from a unique point of view, with a unique set of assumptions, knowledge, and past experiences.

EMPIRICAL ANALYSIS: MEASURES IN PRACTICE

The following chapter is structured in three different sections, dealing with what goes on within the measures, around the measures and how test takers respond to both.

The first section, called 'Normalising potentials,' covers the tests' inherent potentials to normalise test takers and regulate their behaviour or personality. Through value-laden language and colourful charts and diagrams, the measures reveal prescriptive and normative standards and thus encourage test takers to normalise, that is, score within certain numerical spans.

In order for the normalising potentials to be actualised, and to have a performative effect, test practitioners mediate the measures in various ways, as unpacked in the second section, called 'Mediating strategies'. Test practitioners use different strategies to create auras of scientific legitimacy around the tools, and manage test takers' expectations, emotions, and interpretations of their test results.

In the third section, 'Responses to normalising potentials and mediating strategies', I give space to the voice of test takers. I show here the various ways in which test takers respond to both the measures' potentials to normalise and the test practitioners' strategies. Precisely because of the variety of test takers' reactions, test practitioners employ mediating strategies in an attempt to pre-empt resistance and ensure acceptance and conformity. In other words, test practitioners' awareness of the range of possible test reactions motivates their mediating strategies.

This chapter ends with a personal account of being tested and receiving feedback on two different occasions during my study, personalising my object of study and highlighting the peculiar, paradoxical, and emotional aspects of test use.

Normalising potentials

The four measures contain inherent potentials, expressed through value-laden language and graphics, aimed at normalising and regulating test takers' behaviour. The tools express these potentials by considering test takers' scores against predefined desirable quantities of qualities, so to speak, for instance, the appropriate amount of empathy a leader should strive to have (or display). These normalising potentials resemble intentions that are expressed through a somewhat buried authorship of test developers (psychometricians, statisticians, psychologists), granting an authoritative voice to the measures: a voice echoing science and rationality. Through these normalising potentials, the measures (and their developers, sales representatives, purchasers, and implementers) try to influence the test taker to aim for certain scores. By scoring within particular 'spans' or numerical areas, the rationale of the measure and its developers is that a test taker's behaviour is then adjusted to fit the desirable norm.

What the desirable norm then is, varies. At times, the tools promote behaviour or levels of characteristics that fit with the statistical average, that is, what most people in the norm sample scored. Other times, the desirable score, decided specifically for leaders, is to either exceed or score below the average of some particular quality. The measures express these inherent regulative intentions through graphical representations of scores, value-laden language, and other evaluative mechanisms the test reports generate. Thus, the instruments do not simply attempt to measure a person's leadership abilities; they encourage test takers to become a certain type of leader, that is, to aim for a particular score.

One way in which the four measures express normalising potentials is through visual imagery, where gaps, for example between one's actual score and the score to aim for, and high and low scores stand out. In the HD Leadership test report, a wheel resembling a pie chart is used to visually summarise one's scores. The chart consists of colours that correspond with the four main dimensions: personal strength (referred to as 'personal power' only on the summary page, see Figure 3), controller, inspirator, and strategist. An example from my own HD Leadership test results appears as Figure 3.

1. Personal power
2. Controller
3. Inspirator
4. Strategist

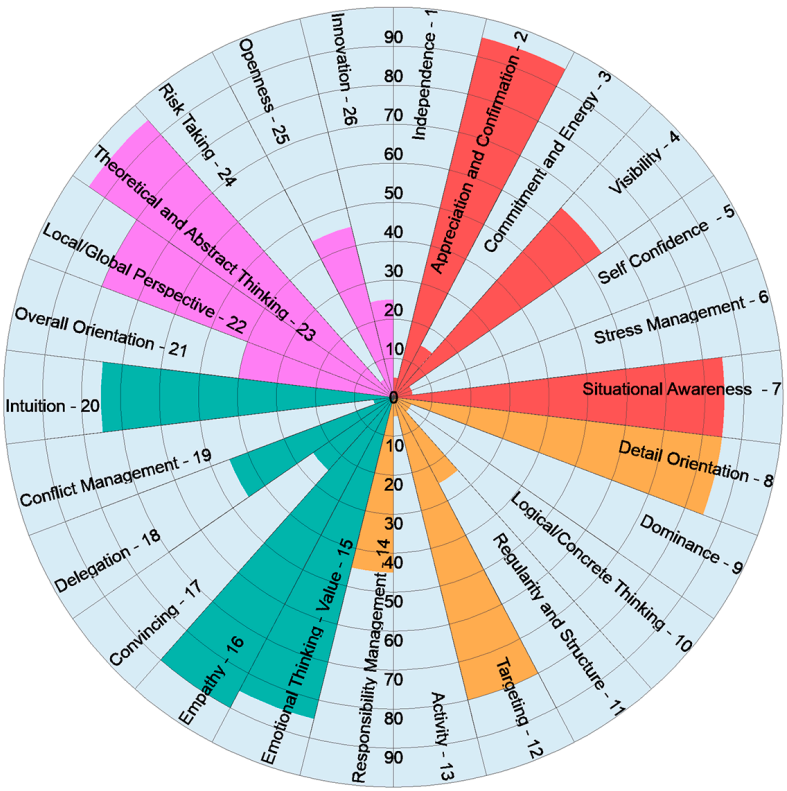


Figure 3: 'Overall summary graph', own HD Leadership report

In this chart both very high and very low scores catch the eye, drawing immediate attention to these. The amount of colour in a field works as a sort of highlighter. For example, 'theoretical and abstract thinking' is almost completely filled, making both the field and the words distinct and noticeable. In turn, 'conflict management' is barely coloured. Further supporting the almost non-existing presence of 'conflict management' is the pie chart image. The design of a pie slice, where one end is narrow and the other wide, means that a very low score is barely

visible. If you look closely, there is a smidgen of green, but from a quick glance, one would probably not notice it. Illustrating one's scores this way thus further diminishes and disregards low scores while strengthening and highlighting high scores.

Also with a purpose of summarising scores and visually presenting an overview, Hogan Leadership Forecast uses a bar chart (see Figure 4):

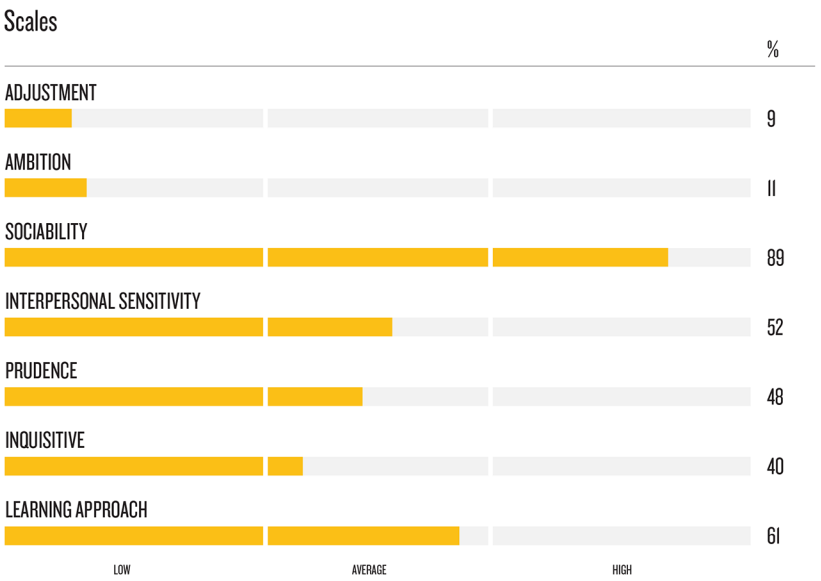


Figure 4: 'Potential' scales, own Hogan Leadership Forecast report

Similarly to the wheel in HD Leadership, high and low scores stand out. Combined with the information at the bottom telling us what scores are 'low', 'average', and 'high', we are already given clues about our level of normalcy. Moreover, the numerical scale format and the coloured bars presented as a summary, without further explanations, make a powerful first impression. In the absence of any nuance or interpretation from a test practitioner, this could lead to exaggerated conclusions (is this person completely maladjusted and unambitious?).

In People Test Person's test report, a quick look at the scale summary also prompts one to notice what scores stand out, such as dots at either end of the spectrum (see Figure 5).

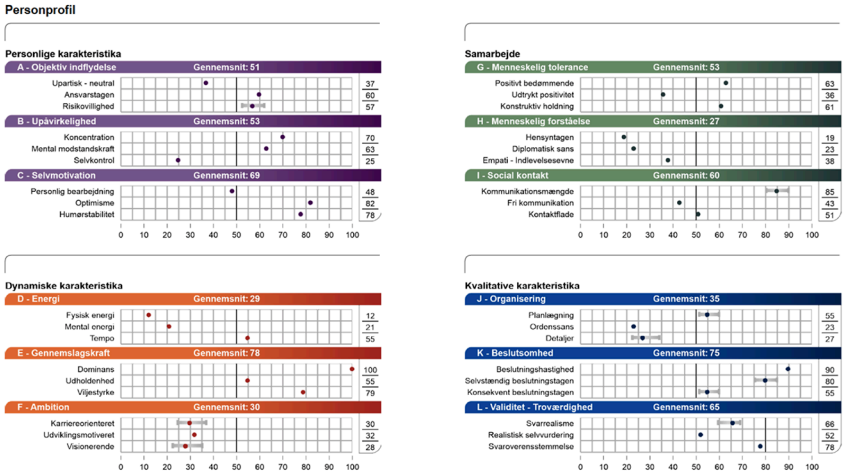


Figure 5: People Test Person sample report

The numbers on the right side further encourage test takers to notice extremes since these give an indication of the level of normalcy one possesses of a certain quality. In this case, 'dominance' and 'physical energy' stand out, since these are the qualities this 'sample person' scores highest and lowest on.

As a 360° tool, The Extraordinary Leader differs from the other three measures in that it explicitly and visually shows the target scores, providing test takers with direct incentives to reach these ideal numbers. According to the instrument's test developers, 'research shows that extraordinary organizational results are the product of leaders who operate in the 90th percentile of competency effectiveness' (Zenger Folkman, online, 2021) and therefore individual scores are compared to the 75th and 90th percentile scores of all 'global leaders'.

Symbol	Norm
	Extraordinary Leader 75th Percentile Norm
	Extraordinary Leader 90th Percentile Norm

Figure 6: 75th and 90th percentile symbols

The symbols shown in Figure 6 are plotted next to all scores (see Figure 7).

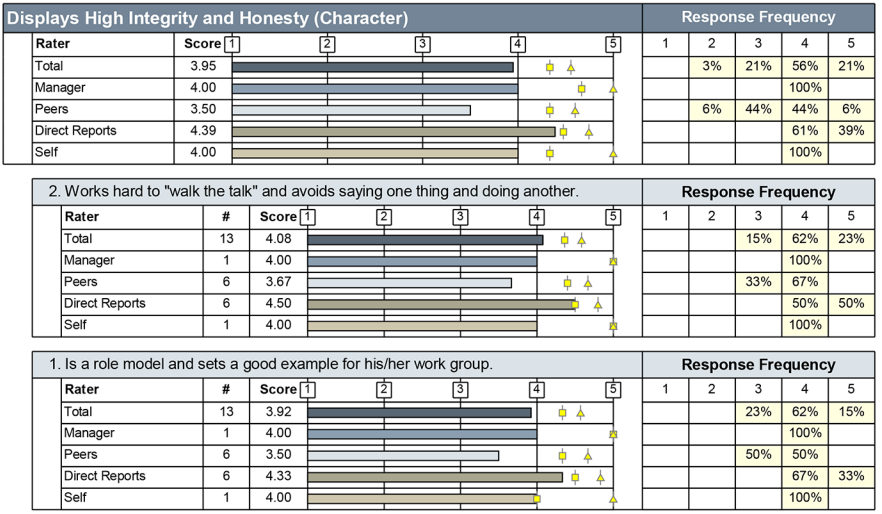


Figure 7: The Extraordinary Leader, example from sample report

Including the 75th and 90th percentile visually cements the gap between a score below these: the larger the gap, the further one is from being extraordinary.

As the name implies, The Extraordinary Leader builds on the idea that a great leader is out of the ordinary, outside the normal and the average. Becoming extraordinary is about 'finding ways to break the mediocrity barrier' (The Extraordinary Leader - Participant Manual, 2015, p.2, module 1). This idea reveals strong rhetorical infusions from charismatic leadership streams (Weber, 1968, 1978). Weber (1968) states for example that the charismatic leader is extraordinary in that 'he [sic] must stand outside the ties of this world, outside of routine occupations' (p.21). Besides being indicated in the name, the book on which the tool is built, *How to be Exceptional,* by Zenger and Folkman, also argues this. An extraordinary leader is, according to Zenger and Folkman, someone who possesses 'a small number of profound strengths that elevate that person above the others' (2012, p.31). Describing the leader as elevated and above others reveals a belief that leaders are extra-ordinary and not part of the 'others'. He or she stands outside the norm.

Signs of the leader as outside or above the normal are also evident in the language of the actual tool. Here, several items emphasise the importance of going 'beyond' the normal 'everyday' work. The test taker is evaluated on whether or not he or she 'has a perspective beyond the "day-to-day" work' and if he/she 'willingly goes

above and beyond'. The leader must exceed expectations and requirements and thus be and act above the norm and the expected. Paradoxically, being able to go above and beyond, being extraordinary, is rated on a set numerical scale from one to five, implying that you can averagely go above and beyond. A set scale cannot be exceeded, meaning that a rating, an assessment of someone will always be restricted to the limits of the scale. Therefore, being elevated and above others arguably cannot be captured on a set scale, a scale that is identical for everyone. A rating of five is still within the boundaries of the scale, which is why even a five cannot be an *extra*-ordinary rating.

Even so, The Extraordinary Leader is designed to direct test takers to exceed the norm, the statistical average. By encouraging scores within the 90th percentile, 'extraordinary' scores will ideally, according to the rationale of the test, become the new normal. The normalising potentials within this 360° tool are thus paradoxical in that test takers are encouraged to score above the norm (normalising extraordinariness), while staying within the set scale.

Besides graphical imagery that draws our attention to (ab)normality and gaps, the four measures use value-laden language that either applaud or problematise certain scores, that is, certain behaviours and personality traits.

After Human Developers presents the test taker's score summary in a pie chart, each scale is explored and the implications of the exact scores are explained. Through mechanisms called 'resource areas associated with a low/high score' and 'areas of potential development' which are only activated or highlighted if a person scores either 'very low' or 'very high' that is, below 20 or above 80 (see Figure 8), Human Developers aims more or less explicitly at normalising their test takers:

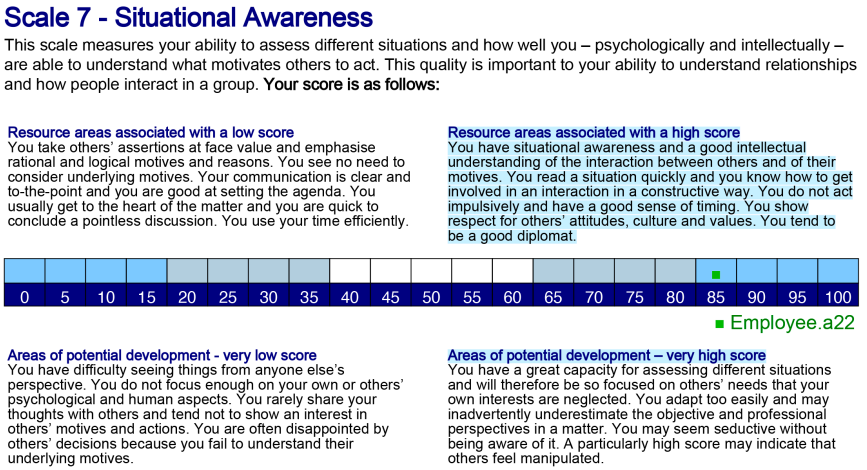


Figure 8: 'Scale 7 – Situational Awareness', own HD Leadership report

Low, middle, and high scores only activate so-called 'resource areas' with mostly positive sounding descriptions, whereas very low scores and very high scores activate 'areas of potential development'. In other words, scoring at both ends of the statistical spectrum (very low or very high) is problematic.

The formulations in 'areas of potential development' are concerned with negative implications of the score, such as: 'You may focus almost exclusively on the overall picture and risk overlooking important details which may turn out to be crucially important,' 'a particularly low score could indicate that you completely overlook important, overall relations,' 'a particularly high score could indicate a lack of ability to work with others,' and 'a particularly high score could indicate that nobody understands what you are saying'.

In the 'resource areas', phrases are more positive: 'You pay attention to detail and are thorough and meticulous in every situation,' and 'you have situational awareness and a good intellectual understanding of the interaction between others and of their motives'. Here, both a low or high score are phrased in positive terms. Only when the scores become very low or high, it becomes potentially problematic for the test takers' leadership practice, according to the test.

Similarly to the HD Leadership test report setup, Hogan Leadership Forecast follows up its scale summary with descriptions of each score. Figure 9 offers an example from my own test:

ADJUSTMENT

Concerns composure, optimism, and stable moods.

BEHAVIORAL IMPLICATIONS

Leaders with similar scores tend to:

- Admit their shortcomings and try to fix them
- Remember their mistakes
- Seem driven and intense
- Take criticism personally
- Have a sense of urgency

LEADERSHIP IMPLICATIONS

Compared to other leaders, your scores suggest that you approach your work with passion and intensity and care deeply about performing well. In addition, you may be easily annoyed with unexpected delays and staff mistakes. On the other hand, you understand when your staff is stressed, you can admit your mistakes, listen to feedback and coaching, and try to improve your performance.

COMPETENCY ANALYSIS

COMPOSURE: You may seem tense or edgy when under pressure, when faced with deadlines, or when others make mistakes, and this, in turn, could affect your team's concentration.

LISTENING: When you are facing deadlines or heavy work pressure, you may tend to stop communicating and listen only for bad news. You can relieve some pressure by planning and delegating before a job starts.

LEARNING AND PERSONAL COACHABILITY: You are open to feedback and interested in improving your performance; however, you may tend to pay more attention to the negative than to the positive feedback.

BUILDING RELATIONSHIPS: Your occasional moodiness, unpredictability, negativism, and tendency to worry can impede your ability to build trusting alliances.

STRESS MANAGEMENT: You tend to be self-critical and intense. You need to learn to be kinder to yourself.

Figure 9: Adjustment score, own Hogan Leadership Forecast report

In the summary, one can read about the implications of one's exact, in this case, low, 'adjustment' score. The value-laden language is softened, and positive spins on the score are included, through phrases such as 'admit their shortcomings and try to fix them', 'work with passion', 'open to feedback'. However, there still appear to be more problematic than good aspects of this score. A scoring of '9' means that, under pressure, I might appear 'tense or edgy', and that I tend to 'stop communicating', 'only listen for bad news', and suffer from overall 'moodiness, unpredictability, negativism, and tendency to worry', which can then hinder my

ability to build relationships. This example is from the report about my 'potentials', where high scores appear to have some advantages and therefore include potentials.

In contrast, in the 'challenges' report, high scores pose a problem. If one is 'excitable,' one is 'overly enthusiastic about people or projects', leading one to become disappointed. The word 'overly' occurs in many of the descriptions. The report on 'challenges' indicates whether you are at low, moderate or high risk of expressing this problematic behaviour. Dimensions at moderate or high risk automatically generate developmental recommendations. These are in all reports written in imperative terms ('don't', 'recognise', 'use', 'think', 'support'). At times the tone is somewhat brusque:

> You probably use displays of emotion as a way of making a point. There are better ways to make a point and repeated emotional outbursts may annoy others.
>
> Practice active listening-don't interrupt.
>
> You need to speak your mind.
>
> Beware of confusing activity with productivity, and don't waste people's time with unnecessary meetings. (Developmental Recommendations, own Hogan Leadership Forecast report)

Other scores appear less problematic; the language is more positive and the recommendations are mostly concerned with doing more of what works. For example, a scoring of 61 on 'learning approach' (in the 'potentials' report) generates mainly positive statements and the recommendations include: 'You tend to take advantage of job-related training and skill development programs. Continue seeking these opportunities', and 'others will normally be able to understand your written memos. Seek feedback on ways to make them even more effective' (own report). Thus, some scores in the Hogan Leadership Forecast prompt more negative feedback than others.

Albeit buried a bit at times, it is clear in the report, that some traits are fundamentally and indisputably good and important to always strive for, regardless of context. For example, a team should be encouraged to be 'creative', one should stay 'current', 'seek feedback', be 'willing to change, 'make decisions', 'learn new management skills', and 'serve as a mentor' (own report). In turn, 'rule breaking', 'inflexibility', 'defensiveness', and forms of reactivity are undesirable traits or behaviours.

In contrast to high and low scores, middle scores sometimes generate statements such as: 'have a normal degree of imagination' (on 'inquisitive'), or 'you are normally cooperative' (on 'interpersonal sensitivity'), suggesting that low or high scores are indeed outside the norm (not just statistically). Again, guiding our attention towards abnormality and normality.

Value-laden language is not only prominent in test reports, it also dominates the educational material I had access to. Human Developers' material even includes a page titled 'The Ideal Manager Profile', where the optimal scores are established, by for example stating: 'no less than 30 in dominance', clearly expressing the assumption or, according to test developers, fact, that managers need at least a score of 30 in dominance. Moreover, the material contains scales with red and green circles, showing where the test practitioner needs to 'pay attention!' (red circle), and what scores are within the 'ideal top leader profile' (green circle) (see Figure 10 below).

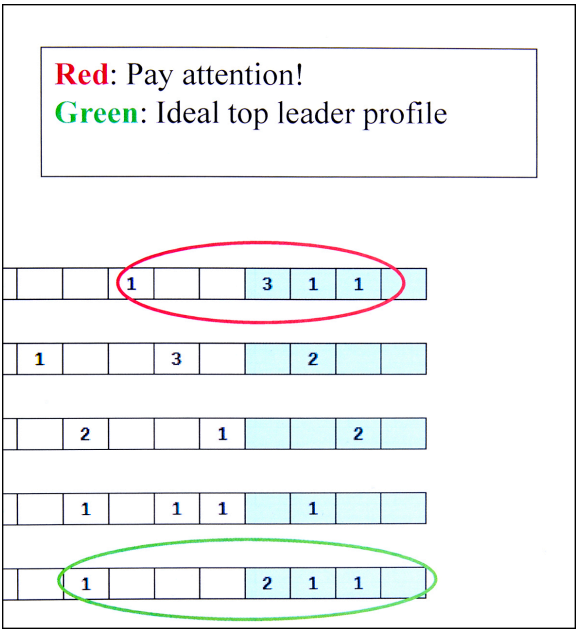


Figure 10: Circles of attention, Human Developers' educational material

This categorisation and value-laden language contradicts the introduction pages to the HD Leadership test report, where it is stated that 'the value of the score is context dependent', as well as their claim in the educational material, where

Human Developers states that the test provides a picture of a person's qualities and dynamics, 'without defining or judging'. Similar caveats are found in the material for the other measures.

In the educational material for People Test Person, assumptions, normative language and taken for granted knowledge dominate. Very high (between 90 and 100) and very low scores (between zero and 29) are described as potentially posing problems. People are expected to do 'too much' or 'too little' of something. In contrast, middle scores 'will rarely cause real problems', and high scores typically just show people's strong sides (People Test Systems, educational material).

Describing one's behaviour as 'too much' or 'too little' implies that the same behaviour is perceived similarly by different people. It further suggests an optimal middle that holds for all people, in all organisations, and in all cultures. Assumptions like these recur throughout the material. For example, people scoring low on 'mental resilience' are predicted to react in ways that others might find surprising: 'You can expect to be surprised over some of the personal reactions that may occur' (People Test Systems, educational material). Inherent in this formulation is the assumption that the people surrounding the test taker will react somewhat identically to the test taker's low mental resilience (score). In other words, that this test taker's behaviour will come as a surprise to the surroundings, implicitly suggesting that others will find his/her reactions emotional or perhaps expressing weakness.

A more explicitly taken for granted phrase is the following: 'too high endurance can be a problem. Here we have the type of person who tends to continue long after people with common sense would stop' (People Test Systems, educational material). 'Common sense' is a translation of the Danish 'sund fornuft'. Both phrases are interesting in this context. The phrase 'common sense' assumes that the knowledge is sound and shared. It also differentiates between the 'common' (which connotes the 'normal') and the 'uncommon' (the 'abnormal' or deviating). In Danish, the word for sense is 'sund' which also means 'healthy' or 'sound', suggesting that people without it are not sensible, and at least in Danish, that they may have an unhealthy approach.

Besides the presence of assumptions and taken for granted notions, many formulations in the material are basically hard to disagree with, for instance that it is good to be responsible, to be able to concentrate, to be more or less stable, or to make an effort.

In People Test Systems' educational material, it is explained that one of the most important learning objectives for future feedback givers is to gain a deep understanding of the different character traits consisting of a blend of qualities. It is also emphasised that the feedback givers learn what hypotheses can be made about the test taker's personality or behaviour, based on the test taker's combinations of character traits. These hypotheses are then supposed to be tested and confirmed as valid (or not) in the feedback session. People Test Systems provides a list of examples of these:

> The (too) kind person, the (too) dominant person, the temperamental person, the impatient person, the perfectionist person, the (over-) ambitious person, the vulnerable person (wearing velvet gloves), the person with low self-esteem, the "administrative disaster", the misjudging person. (People Test Systems, educational material)

This is followed by disclaimers and softening formulations like the somewhat ironic phrase: 'It is important to stress how nobody likes to be put in a box'. Arguably, the typification above functions to categorise people, which indeed resembles putting people in boxes. It is emphasised in the educational material that these descriptions are in fact hypotheses, that test takers should be met with open, curious questions, and that the different characteristics should be seen as part of a whole. However, such disclaimers and softening sentences are only necessary due to the value-laden language in the test. If the language was less categorical, it would not be necessary to make such qualifying statements to lessen its impact in the first place.

In the educational material for People Test Systems, certain values are consistently attached to particular scores. A high score on a characteristic is generally positive. This is in clear contrast to the type of descriptions for low, very low, or very high scores. The following examples indicate the use of value-laden words to describe different levels of 'self-control' and 'mental resilience' (emphases added):

> **Characteristic: Self-control**
>
> High level: Here is a person who is *well-balanced*, who seems to have a lot of *self-control*. The person will often have a *soothing effect* on others and exude *calm.*
>
> Low/very low: Here is a person, who is experienced as *temperamental* or who easily gets *nervous* and *insecure/uncertain*.

Very high: The person is *so controlled, that it doesn't seem natural*/that it seems *forced*. The person is *hard to read*. Doesn't react until everything is *chaos* in which case, the reaction might be surprising.

Characteristic: Mental resilience

High level: It takes a lot for this person to be affected personally. The person is *mentally resilient* and doesn't get too affected by the things life and work throw at them.

Low/very low: *Touchy* and *vulnerable*. *Even little things will be taken personally*, and you can expect to be surprised over some of the personal reactions that may occur.

Very high: *Doesn't fear to be affected* at all. Therefore, the person can also have *trouble empathising* with others who more easily get affected, resulting in the person seeming *tough*. The property can also mean that the person simply continues his/her activities completely *unconcerned with previous mistakes or failures.* (Educational material, People Test Systems)

It does not require much interpretation to infer that the scores to aim for here are high scores. The two extremes, very low and very high, are described in a negative tone with emphasis on the problematic aspects of the expected behaviour. What is of interest is again that this is in contrast to the preliminary introduction where test practitioners often stress that there are no good or bad profiles.

In sum, the four measures have in common a way of promoting ideas of normalcy and desirable leadership conduct. The measures are infused with different kinds of normative assumptions and are evaluative both in language and graphical representations, generating prescriptive conclusions. The visual representation of one's scores allows for quick comparisons, assessments, and identifications of deviations, and ultimately evaluations of one's level of normalcy. Whether presented on a colourful wheel, through bars, or dots on a spectrum, high and low scores stand out. This makes it difficult not to pay attention and attach value to such scores. In addition, the value-laden language in both the test reports and educational material reveals that some scores or combination of scores are expected to lead to problematic behaviour; problematic as assumed in the specific test with its particular understanding of what constitutes good leadership.

Confronted with these tools' value-laden prescriptions, test takers respond in various, unpredictable ways, ways that test practitioners aim to either pre-empt, moderate, or support. In relation to this, test practitioners work to foreground or downplay the normalising potentials in the measures. Overall, they frame the instruments by using particular strategies to mediate test takers' or future feedback

givers' experience with the instruments. Ultimately, test practitioners seek to convince test takers of the measures' value and legitimacy. How test practitioners carry out these efforts will be unfolded in the following.

Mediating strategies

In this section, the context of the four measures and their normalising potentials takes a central role. The context includes the social actors framing the measures and the spaces where these framing efforts take place. This, since it is not enough to examine the measures out of context; in order to understand how the measures operate in practice and what informs their use and potential effects, we must also explore how social actors mobilise soft norms around the measures, and overall try to actualise the normalising potentials within the measures. Such a perspective provides insight into the way quantitative tools do not work by themselves, but are framed by test practitioners in the anticipation of the variety of (un)predictable ways test takers respond to being tested.

The first time I experienced a test representative in action was when an external consultant introduced The Extraordinary Leader to test takers at PharmExtra. In the following, I draw on some of my field notes to convey my experience when I entered the field.

When I arrived at PharmExtra on April 30th 2018 at 8am, Aaron, my contact person and the company's Leadership Specialist, met me outside the building and then took me to the site of the workshop. We went through an open office space where the Training and Leadership Development department was located. A combination of clear and frosted glass separated the open office space from smaller conference rooms. In the frosted glass, words were written in horizontal lines, such as: 'COMPETETIVE TRANSPARENT FLOW MINDSET', 'INNOVATION RESPECT DECISION MAKING', and 'EXCELLENCE MAXIMISE SKILLS'.

I was led into a small room with a raised computer and a headset, which immediately caused some confusion since I had expected a workshop where participants were physically present. Without inquiring about this, I tried to infer from what Aaron said how the workshop would take place. It turned out the workshop was virtual (through Adobe Connect): All 36 participants (two groups, one in the morning and one in the afternoon) joined the workshop via their computer, meaning that they could participate irrespective of their location

(whether they were located in Denmark, Brazil, U.S., or France), also suggesting to me that the workshop would be held in English. After a quick mental evaluation of the pros and cons of this where I thought about the limitations in terms of interacting with the participants, I focused on the positive. This virtual setup would allow me to systematise my observations more easily and less noticeably. I could take notes, pictures, and record impressions without disturbing or affecting the participants.

After being introduced to the external consultant, Michael, who sat in the adjacent room with a technical supporter, I prepared for the commencement of the workshop. A countdown clock appeared with big red numbers, reminding me of a bomb. I shared this thought with Aaron, who replied: 'Well. It sort of is like a bomb'. At 00.00.00, the workshop was in progress. Aaron ran some initial technical checks, followed by a short talk about conduct where Aaron encouraged the participants to be active and participate. Then Michael took over and introduced himself. Among other things, he mentioned his background in psychology and finance. His English was almost flawless. It turned out he had worked in investment banking in London for 12 years. Michael exuded experience, confidence, and trustworthiness. He delivered his points in a calm, clear, well-articulated, and organised way.

My second experience with a test facilitator representing a tool, was at the certification workshop at People Test Systems.

I arrived at People Test Systems on September 18th 2019 at 8am. It was my second time there. The first time involved a meeting with Emma and Elizabeth about my project, expectations, and the possibility of attending the workshop. I therefore knew how to find the right floor, which lowered the nervousness a bit.

Before arriving at the certification workshop, I had received a link to some e-learning material about People Test Person that the other participants had also received. The material was structured in eight modules concerning different aspects of the test. It included its 'construction and background', different characteristics: 'personal', 'dynamic', and 'qualitative', ethical guidelines, and questioning techniques. The e-learning finished with an online exam that one had to pass in order to attend the certification and become an authorised test practitioner.

I took the test myself and found it surprisingly difficult, especially the part where I had to choose what combinations of characteristics would lead to what kind of behaviour. Combinations of characteristics I intuitively figured would lead to a

certain type of behaviour would turn out to be wrong, more often than not. This experience made me realise the importance of learning the test language, so to speak. What might intuitively seem to mean one thing may mean something very different in a test context. However, I might have spent more time with the material if my participation relied on getting a pass. Even though there was no pressure on me to pass the exam, no instrumental gain from passing, I still felt the disappointment of failing it. I wondered how the atmosphere at the workshop would be and if I would be able to understand the content at all.

A woman that I had met the first time I visited People Test Systems whose name I recognised from the email with the e-learning link, showed me to the room where the certification would take place. The instructor, Karen, and a couple of participants were already there. Polite head nods were exchanged. The participants sat on their designated chairs around a table. I quickly found my seat, up front, next to the instructor. On the table were fruit, snacks, coffee, and tea. The ceramic mugs were surprisingly nice, nothing like the standard white porcelain mugs at workplaces in the public sector where I had previously worked. The snack and tea selections were also impressive, all in all giving the workshop a slightly luxurious feel. It suggested, to me at least, that People Test Systems can afford (or wants to signal that they can afford) such a level of comfort because their product is successful. With an appealing first impression, and by feeling pampered, participants might be more inclined to trust their product and advice.

In front of each participant's spot around the table was a programme revealing the plan of the two following days. The programme was quite packed: it started at 8.30am and finished at 16.30am with only one 30-minute break for lunch, perhaps signalling that the content is rich and demanding because the product is complicated and 'dense'. Next to the programme was a list of participants: 11, including myself.

Aware that the other participants probably assumed I was there to gain certification, I tried to imagine I was in that situation. This led me to read the programme differently, noticing every time the word 'exam' appeared. I immediately felt at ease reminding myself that I was not there to pass an exam.

The instructor, Karen, and I quickly talked before the workshop started. She encouraged me to emphasise to the participants that my purpose was not to evaluate their performance in any way. Emma and Elizabeth had asked me to do the same. I reassured her about my role there, and that I would try to appear the least 'evaluative' possible.

The workshop began with a round of introductions. Karen asked me to start, perhaps to disclose my role at the outset. I stated my name, place of work, why I was interested in attending this workshop, and how grateful I was to be allowed. I mentioned several times that I was not there to evaluate them as future test practitioners, but that I was simply interested in how practitioners were certified in a test such as People Test Person e.g. how the theory the test relies on was communicated to them. The participants all appeared comfortable with this. They smiled and did not ask me any questions.

During the workshop, my note-taking did not seem to cause any distraction or prompt curiosity since several of the other participants did the same. I simply looked like just another participant eager to scribble down Karen's points which she conveyed very enthusiastically. She stood up, gesticulating passionately with her eyes wide open. She spoke so loudly that I almost instinctively moved my chair slightly backwards. Granted, I was sitting very close to her, but the room was quite small with only 11 participants. Whenever someone commented on something, Karen replied with a loud 'yes!', 'very good!', 'exactly!', or 'agreed!'. Similarly to Michael, Karen delivered her points in an organised, confident, and almost proud way.

Even though the measures are constructed and designed in a way that is meant to give them a serious, reliable, and scientific feel, test practitioners still employ different activities and mechanisms to further create the impression of scientific legitimacy. Through certain ways of introducing and framing the tools, giving test feedback, and certifying future practitioners, consultants influence the uses and effects of measures. In the following, I discuss examples of such influences and present five mediating strategies.

Mediating strategy #1: Creating legitimacy and trust

At the certification workshop at People Test Systems, the instructor, Karen, repeatedly told the participants how valid People Test Person was, emphasising its trustworthiness, that it had been designed and constructed in a 'solid way' and had been 'investigated'. She argued tautologically: 'When we measure dominance, then we measure dominance, because we have a whole team of scientists making sure of this' (dominance being one of the so-called 'dynamic characteristics' that are measured in the test). Scratching the scientific surface, she stated it is 'based on correlation theory and other mathematical stuff', without going into more detail.

In her view, it is not up to practitioners to explain this scientific basis in detail. If test takers question the test design and how constructs are measured, Karen urged the future feedback givers to avoid giving long explanations and simply say 'we know. We know that the construction is solid'. This claim was justified by referring to the 'wise people' behind the test design: 'They are psychometricians, they know things about psychology and math, so they know how to calculate things' (Karen, certification workshop).

Test developers and practitioners generally tend to frame chosen characteristics, such as dominance, as self-evident, factual, and unproblematic. Particularly in test materials, characteristics are presented as indisputable, self-explanatory, and of obvious importance to the user.

My experience at PharmExtra was similar. I witnessed the time and effort a test practitioner employed to frame The Extraordinary Leader as legitimate and trustworthy.

After Michael, the consultant, had introduced himself to the test takers, he laid out arguments supporting the 360° tool, presented as different 'insights'. Michael talked about the benefits of the tool and the 'evidence' supporting it. He emphasised several times that there is: 'data supporting causal effect', 'strong and linear correlation', 'fantastic effect', 'radical impact'. He stated that: 'the better/the higher you score, the greater impact', 'I had prejudices, but then I learned', 'self-perceptions are not as accurate as other's perception', and 'a 360 is just a great thing to get'.

Along with these statements, Michael presented data (as numbers and graphs) supporting his arguments. He used this, for instance, to quantify great leaders' impact on customer satisfaction, income, employee engagement and the effects which 'profound' strengths (competencies at the 90th percentile) have on leadership effectiveness. Numbers played a heavy role throughout the session and were used as the apparent primary means to support Michael's arguments and assure the audience of the tool's legitimacy and reliable scientific foundation.

Michael engaged the participants by using different features. The participants answered polls on 'which insight did you find most interesting?', and 'what is the impact of great leadership on business results?' These polls reinforced the points made by Michael. By making the participants phrase the insights and choose the 'correct' option themselves, Michael made sure the participants understood, appreciated, or realised his arguments and stated beliefs. Using polls might also create feelings of participation, influence, and autonomy. Michael invites the

participants to voice their reflections, reducing a possible feeling of being forced, but without having any actual influence over the process.

Another feature supporting the quest for buy-in was the monitoring of people's attention to the workshop. Next to the participants' names were bars in a colour that revealed their level of engagement (Figure 11).

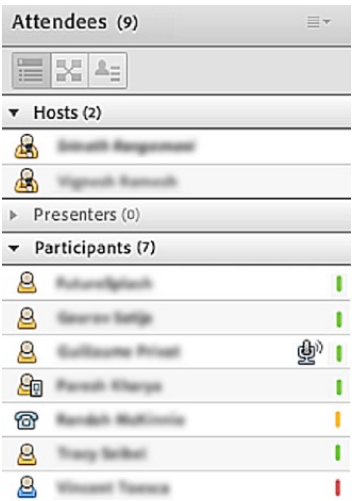


Figure 11: List of participants and their engagement level, virtual workshop

Adobe Systems Incorporated, the developers behind the computer programme, explains how participants' engagement level is calculated:

> Different activities have different engagement point values. The following are events and actions that provide engagement points to an individual:
>
> 1. Explicit activities give 100 points to an individual including:
>
> a. Chat activity – Chatting privately and publicly
>
> b. Q&A activity – Asking questions, adding, assigning, and deleting
>
> c. Poll activity – Responding to a poll
>
> d. File download – Downloading a file
>
> e. Notes pod activity – Typing
>
> f. Status updates

i. 10 points if status is stepped away

ii. 100 points for all other status updates

g. Mouse activities – 80 points for activities like clicking to start a webcam, scrolling a notes pod, and so on

2. Meeting window focus

a. 70 points if meeting browser tab/add-in has focus

i. otherwise, 20 points

b. 10 points if meeting add-in is minimized (Adobe Connect User Community, 2023, n.p.)

Monitoring participants' level of engagement was thus an in-built feature in the program PharmExtra used for their virtual leadership development. The participants' level of engagement (as measured by the programme) became in this way very visual. If the workshop had been physical, one would have had to evaluate the participants' engagement in other ways. For example, are people slouching, yawning, glancing at their phones under the table, or are they taking notes, looking at the consultant, perhaps nodding or otherwise responding to the consultant's arguments and points. In a virtual format, we could simply speculate that some participants were less engaged than others, since they had perhaps minimised the computer window or refrained from answering one of the polls.

By engaging the participants in different ways (and monitoring this), the consultant attempted to convince them that the tool is helpful, reliable, and valid. Reactions both during and after the workshop show that this work had some immediate effect.

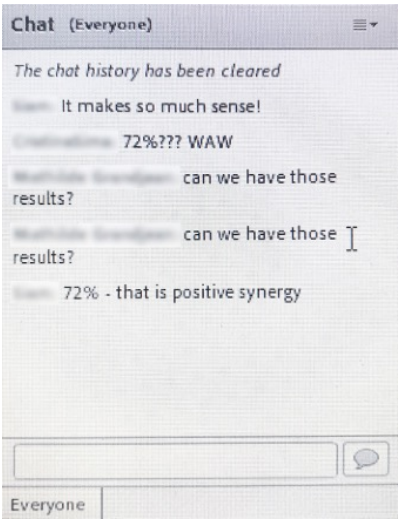


Figure 12: Chat window, virtual workshop

Figure 12 shows a picture of the chat window in which the participants could write questions and comments throughout the workshop. In this particular picture, we see different participants' reaction to Michael's assertion that there is 72% likelihood of someone becoming an extraordinary leader if this person combines the strengths 'builds relationships' and 'drives for results' (based on some mathematical model). Interestingly, two of the participants react specifically to the number '72'. This number sparks an enthusiastic 'WAW' [wow] and '- that is positive synergy'. This is an example of how numbers are used, here by Michael, as an effective way to create trust, legitimacy, and overall belief in the tool. How this particular number was arrived at is not explained to the test takers. Based on the test takers' reactions however, the number does not seem to be in need of further justification, it carries and conveys an authority in itself.

Throughout the session, there were some group work exercises taking place in smaller breakout rooms. Participants were digitally allocated into groups, with their own virtual forum, where they could talk freely and not listen to the other groups' discussions. I was able to listen in on these breakout sessions. At one point, I actually felt like a fly on the wall. My presence was almost unnoticeable, since my microphone was turned off and the only sign of my presence was the addition of my name to the list of participants in the group.

At one of these breakout sessions the participants analysed a sample test profile. They talked about how this person 'lacked' certain competencies, and was 'better than the norm, but maybe not extraordinary', i.e. 'not really a superstar', why they recommended that he got a 'distinct profile that he [could] promote'. Not only was the language clearly affected by the terminology presented to the participants just beforehand, they also quickly adopted the entire framework about distinctive strengths and the importance of being extraordinary.

At the first session, the participants had not yet received their 360° reports. They had completed them, so they knew what they were measured on, but they had not yet seen their results. Before the second session, three days later, the test takers would receive their reports. I wondered if withholding the 360° reports before the first session was deliberate; a way of avoiding or reducing potential criticism. Before the test process would become too personal to the participants, they might be more inclined to listen to all the arguments about the helpfulness of the tool. Telling the test takers about the value of the tool, before they receive their results, might also be a way of 'softening the blow', so to speak.

Michael's efforts to make test takers trust the tool continued after the workshop. He offered all participants a one-on-one coaching session where they could discuss their 360° results, e.g. how to interpret different ratings, and what to focus on in the future. At Joseph's coaching session on Skype, Michael pointed out and argued for the high validity of the report, based on the response rate (13 out of 14 people had agreed to rate Joseph on the 49 items). Michael therefore encouraged Joseph to trust or embrace the results, based on the fact that the generated report gave a 'good, representative image'.

The test practitioners' creation of trust and buy-in is further supported by test purchasers' preconceptions about the value of quantitative assessment tools, making the sale easier. The Programme Director at BigBank, Megan, who had chosen to implement Hogan Leadership Forecast, states:

> Models and tests can create the foundation for a dialogue where you kind of avoid "this and that person thinks that Megan is really weird" or "it is really annoying that", to "oh right, you have a preference for…" then we can sort of lift it up and away from me personally and into a place where we have a shared language or vocabulary: "Ah, so you don't think in that way." (Megan)

Tests appear to answer Megan's call for a distanced, more detached perspective that replaces personal opinions.

Concerned with his organisation's legitimacy, Aaron from PharmExtra, responsible for choosing The Extraordinary Leader, points to the benefits of using a valid tool, represented by a trustworthy consultant:

> I think that it is becoming more and more the case today that you have to be able to prove your views and we feel that Zenger Folkman knows that they have this big database which the recommendations of what cross-competence fits your strengths is based on, so that means a lot and we can feel that in the participants as well. They also emphasise that "okay there is actually something about this" and they are happy about it too. (Aaron)

According to Aaron's reasoning, in order for the Training and Leadership Development department to offer legitimate, substantiated views and assessments to the rest of the organisation, they need a trustworthy tool, which has led them to choose The Extraordinary Leader.

The simple numerical format of the tests prompts positive expectations of their trustworthiness and overall value, inclining test-buyers to trust the tools.

Test practitioners' efforts at test introductions, workshops and certification programmes to create legitimacy and trust in their tools mirror and further build on how the measures are presented on their websites.

For example, on Zenger Folkman's website they claim to be 'globally recognized as the most efficient and effective 360' (Zenger Folkman, online, n.p.) and describe their solutions as a 'proven approach' that brings 'science to the art of leadership', by putting 'leadership under the microscope' (Zenger Folkman, online, n.p.). Words like 'proven', 'science', and 'microscope' initiate associations with laboratories, mathematical proofs, and physical facts. Exactly what Zenger Folkman is inspired by and strive for:

> Consider the progress made by the medical profession as they have embraced the concept of their practice being strongly guided by rigorous scientific evidence. Frankly, with only a few exceptions, such rigor has been lacking in the study of leadership. (Zenger, Folkman & Sandholtz, 2021, p.2)

A proven, scientific approach is implicitly put in contrast to what the test developers might regard as the oppositive, such as unsubstantiated opinions. Choosing an approach that is proven and scientific, one can blindly trust and vouch for the outcome.

Similarly, Hogan Assessments establishes the legitimacy of their tool by referring to its scientific validity and consequently its trustworthiness. One of their main headings on their website reads: 'Turn leadership into an exact science' (Hogan Assessments, online, n.p.), suggesting that with Hogan Leadership Forecast, one can obtain exact and scientific knowledge. Implicit in this statement is that without this tool, one must settle for inaccurate and unproved knowledge. According to Hogan Assessments, settling for that both can and should be avoided, since: 'Personality can be assessed with accuracy and reliability' and 'personality assessments compare people *objectively*, on equal playing fields' (Hogan Assessments, online, n.p., emphasis added). Besides providing an objective instrument, Hogan Assessments claims to offer a tool that '*truly* differentiates high performers from their peers' (Hogan Assessments, online, n.p.), presupposing that the organisation has already identified their high performers, and that it is of great value to 'truly' know how these high performers differ from their peers. The latter assuming that the test offers the capability to provide truthful knowledge.

The measures are further legitimised in educational material. Human Developers emphasises in their certification material that 'the questionnaire has turned out to be very robust' and that the tool is 'efficient, thorough and accurate'. The test developers stress how their tool offers a 'precise x-ray' of the measured person, suggesting that a person's inner personality factors can be known in the same way an x-ray reveals bones and muscle. When Human Developers uses this particular analogy it creates associations to physical facts, prompting the test user to expect accuracy and indisputability of the test results. Moreover, an x-ray separates the measurer from their object, implying that it is indeed objective.

In sum, test practitioners work to present their instruments as devices that accurately measure one's reputation, behaviour, preferences, or personality, which are crucial for assessing one's leadership style or effectiveness. When one then encounters the measures' normalising potentials, these appear scientific, well-documented and more or less beyond question.

Mediating strategy #2: Managing expectations

As part of framing the tools in certain ways, test practitioners manage test takers' expectations in different ways. On a general note, practitioners talk about the tools in particular value-neutral ways, supporting a positive, non-instrumental purpose

and representing a form of 'front stage talk'. This appears to be used to lower the stakes and reduce any fear or resistance towards the tools.

Moreover, Michael, the consultant for The Extraordinary Leader, managed the participants' expectations at the workshops by using different techniques that together construct a sense of urgency and importance, establishing acceptance as the norm, the emotional ideal, and feedback as a 'gift'. Similarly, Karen, the instructor at the People Test Person certification workshop, attempted to regulate future test takers' expectations through particular ways of training future feedback givers.

Test practitioners' front stage talk about the measures was used in interviews with me and on two occasions of observation at BigBank and People Test Systems, respectively. In order to create the desired image of the measures, test practitioners talk about the tools in value-neutral terms, emphasising the tools' soft and non-instrumental purposes.

The measures are described as primarily personal and meant for personal development (Julie, Megan, Aaron), they are supposed to create insight (Elizabeth), dialogue and shared language (Jacob, Megan, Julie), and reflection (Violet, Eva, Megan). They legitimise conversations and strengthen relationships (Julie), provide feedback on how one is perceived from the 'outside' (Aaron, Violet, Elizabeth), and are an opportunity to evaluate oneself (Violet). They target development and identify strengths (Aaron) and potentials (Julie).

At the certification workshop at People Test Systems, Karen, the instructor, emphasised several times that People Test Person is not a test, there are no right or wrong answers, and low or high scores do not imply positive or negative personality aspects. As mentioned earlier, the same is written in the material for HD Leadership and Hogan Leadership Forecast.

This was met with some discussion. For instance, a certification workshop participant remarked: 'When I had a low score, I was thinking "Oh no", because a high score should be good', to which Karen, replied: 'It is crucial to stress that it is not important if it is a high or low score'. Karen was generally careful to communicate that low scores are not bad, by for example directly encouraging the participants to not 'value the scores – we shouldn't start saying "that's good", "that's great", "that's okay", "that's not great"'.

This front stage talk contradicts the fact that a person cannot score lower than five, since 'an empathy score at zero – no one wants to have that. You can't have

a score lower than five. The minimum is five. It's too harsh with a zero on "mental resilience" or "self-control"' (Karen). Clearly, some scores are indeed 'not great'.

At the community meeting at BigBank, one participant asked why the bank had chosen to implement Hogan Leadership Forecast and measure people in this way, to which BigBank's Transformation Consultant responded: 'We have seen that the language, the terminology has been very beneficial … The primary purpose is to create a common frame of reference … that you have the same starting point'. Megan, the Programme Director, added to this and stressed the importance of having 'something' that supports test takers' own individual development: 'It is a tool of support … it is an individual report, no one else is getting it but you. For your reflection on where to develop … no one is good or bad, we all bring our own personality to work, we need to conduct authentic leadership' (Megan). Megan was careful to present Hogan Leadership Forecast as harmless and in the test takers' own personal interest, working to manage the test takers' expectations and attitudes towards their forthcoming test experience.

Likewise, at Logistica, Julie, who facilitates HD Leadership, emphasised that the tool should not be used for any value-laden purposes:

> In the feedback there is a risk that one might not be skilled enough to convey that high does not equal good and low does not equal bad, and there are no right and wrong profiles, la la la. And they are not a score. And it is my general opinion on tests that it can be a bit delicate, someone can be afraid that this is a measurement instrument. That one can get a good test result and a bad test result. (Julie)

According to Julie, there are no good or bad test profiles. Instead, she explained, the purpose of testing is to identify and cultivate potential.

In sum, whenever test practitioners are met with questions about the measures' purpose, they tend to respond in ways that emphasise the 'softer', 'non-threatening' aspects of the tools. Besides talking about the measures in non-instrumental terms, test practitioners also employ more concrete tools to manage test takers' expectations.

The first part of the virtual workshop for The Extraordinary Leader was about the impact of 'great leaders'. This functioned as a persuasion strategy to convince the participants of the value of the tool, but it also worked to manage test takers' expectations of themselves. Statements like: 'Good does not equal extraordinary – and your organization needs you to be extraordinary' (PowerPoint slide from the workshop) both create a sense of urgency while appealing to the participants'

sense of responsibility and desire to be important for the organisation's success. Implicitly, the statement conveys that the test takers should want to reach this extraordinary state, if not for their own sake, then for that of their organisation. The organisation relies on the participants to develop as leaders, potentially influencing their expectations of themselves and willingness to, at least overtly, embrace their results.

At the end of the first workshop session, after having established the legitimacy of The Extraordinary Leader, Michael, the consultant, presented the participants with different emotional stages that 'most experience to feedback.' This was alluded to by the acronym 'SARA' (surprise/shock, anger/anxiety, rejection/rationalisations, and acceptance). First, Michael revealed that 'I definitely went through this', making the reactions more credible and perhaps inevitable. Then, he introduced the stages by asking the participants to guess what each letter stood for. At the end, he told them that he often experiences people being 'stuck' in anger and 'especially rationalisations'. He mentioned that 'rationalisations are normal, but I urge you to move away from rationalisations and on to acceptance'. Michael then directly encouraged the participants to reach 'acceptance' as quickly as possible.

This way of guiding test takers in the direction of acceptance is supported by the first pages of the 360° report, where one is met with the following sentences: 'As you review this report, keep in mind that feedback is meant to be constructive. You will derive the most benefit from it if you keep an open mind, rather than becoming defensive or looking for reasons why it "must be wrong."' Having an 'open mind' is put in direct contrast to being 'defensive', a word with negative connotations. Constructing this opposition limits the field of possible attitudes: You can either be open-minded and derive benefit, or defensive and gain nothing helpful from the report.

As part of preparing the participants for their test reactions and emotions which might include being 'stuck' in defensiveness, the participants received a 'Top ten list of rationalizations' (Figure 13).

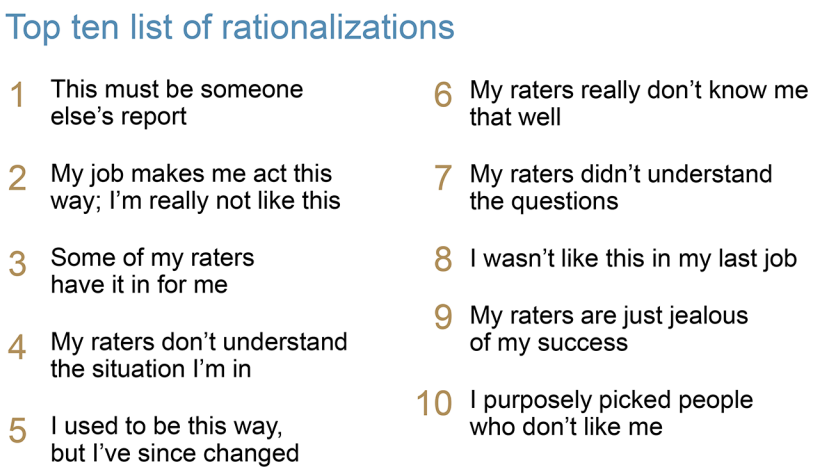


Figure 13: 'Top ten list of rationalizations', PowerPoint-slide

Michael ridiculed these, indicating that they were 'silly' to experience and perhaps restricting the test takers from expressing these reactions. The rationalisations were established as 'typical', something Michael had seen many times before, making the participants' reactions less personal, unique, or even that serious. Most importantly, the reactions were presented as simply a step on the way to an inevitable and desirable acceptance.

Besides constructing acceptance as the norm and the appropriate reaction, Michael repeatedly mentioned the slogan-like statement, 'feedback is a gift'. This was also used as one of the slide titles on the PowerPoint show, about which Michael explained:

> In my experience, receiving a 360 is always a really, really interesting thing. And also, it can be a little bit tough to receive a 360, whether you have received it before or not. But there are definitely always benefits in receiving it. The first thing to know about a 360 and feedback in general is that feedback is a gift. I know this is a huge cliché, but it is actually still the truth … it is like with Christmas Eve. Sometimes we get a gift and it's not exactly what we had hoped it would be. And sometimes that's the case with feedback as well. Sometimes it surprises us, maybe it is better than we expected, sometimes it's different or worse than we expected. But it's generally always a gift. People have taken, each of your raters has spent at least 20 minutes giving you feedback, honest feedback, and generally that's a really

> good thing. And you should assume genuine intentions … please assume that the raters have genuine intentions. (Michael)

Here, Michael speaks to the test takers' emotions. He acknowledges and emphasises how the 'feedback gift' might not be on top of the wish list but repeatedly encourages the test takers to still consider it a valuable gift. Michael compares receiving feedback with accepting an ugly sweater 'like a gift from a mother-in-law', playing on feelings of courtesy.

Later, the importance of gratitude is further cemented, when Michael says: 'I would definitely encourage you to thank the people who filled them out and to share the results with your manager'. Convincing the test takers that feedback is a valuable gift encourages them to accept and embrace their results. Also, emphasising the time and 'genuine' and 'honest' effort spent by the test takers' raters is a way of instilling gratitude and humility in the participants – urging them to accept and appreciate their feedback.

Aaron, the Leadership Development Specialist at PharmExtra, also emphasised that feedback is a gift: 'It might not be a great experience for the participant, but it must be a nice gift even though it is negative, to find out how people see me, because then I can change it'. Both Michael and Aaron made the point that feedback is a gift even when it might be negative. The risk exists that negative feedback may be discarded or rationalised away. In contrast, by establishing that 'feedback is a gift', the participants are steered into its acceptance.

A quick shift to the perspective of test takers shows us that some indeed appeared to adopt this view: 'For me it has been a gift to do this 360 degree' (Carl), 'for me it was a gift, a really great tool for development' (Oliver), and 'I take it 100 percent in, like a gift, now that people have told me something' (Richard).

Working to establish the gift of feedback and its inevitable acceptance, Michael prompts certain expectations the test takers should have of themselves, their reactions, and the tool: the test takers should be grateful and accept their results.

At the certification programme at People Test Systems, expectations were likewise an implicit central theme. Here, Karen, the instructor, indirectly regulated future hypothetical, test takers' expectations by training practitioners to give feedback in a certain way. This involved, in part, creating a comfortable feedback situation. Karen asked participants to discuss in groups what elements are important to consider when giving good feedback. Participants mentioned making the test taker feel safe, not putting people in boxes, appearing calm, controlling body

language, emphasising that the tool is trustworthy and that it is not meant to sort people, and that there are no right or wrong answers.

Later, Karen handed the participants a manuscript they were encouraged to use, with the following instruction: 'There are some things we want you to say'. This consisted of eight points the participants should follow chronologically when giving feedback. It would begin with welcoming the test taker, asking them about their previous and current test experience. This should be followed by a presentation of the conversation to come, for example going through the purpose of the test. After this, the feedback giver is supposed to clarify and align expectations and then present the actual analysis. In presenting the analysis, feedback givers are encouraged to mention some background to the test, for instance that it is normative, not ipsative, that it measures behaviour in 33 areas, that test takers are compared to a norm group of 3000 people, that it is not quality that is measured, but the amount of a particular behaviour compared to the norm group. After this, the parameter of trustworthiness is to be explained, followed by how the feedback giver will give the feedback.

At this point, the actual feedback begins, which is described as an exchange between interview and feedback. When this is done, the feedback giver is supposed to summarise strengths, challenges, and areas of development, ending with asking the test taker for any questions. In addition to this, at the bottom of the manuscript, three bullets appear under a headline reading 'important':

> To create a good connection
>
> To create trust
>
> To create a good atmosphere

In Human Developers' educational material, future certified test practitioners are similarly urged to give feedback in a certain way. The authors of the material state that 'feedback must at times overcome great resistance' and 'it is necessary that the consultant applies a very specific method when he/she gives feedback'. More specifically, the authors encourage consultants to always be 'appreciative in approach', especially if the test taker has many low or very low scores.

Emphasising the importance of a good connection, atmosphere, an appreciative approach, and trust suggests that test practitioners are aware that test takers may not perceive these tools as non-threatening, harmless, and neutral, even if test practitioners present them as such. The measures require the right framing, such

as a certain procedure for giving feedback, in order to adjust test takers' expectations and responses and thereby reduce resistance.

In sum, test practitioners preparing (guiding) test takers' future reactions, providing them with a certain terminology, certifying practitioners in ways that encourage them to give feedback in a specific way, are examples of test practitioners' attempts to manage both practitioners' and test takers' expectations and as a result, willingness to receive their results with openness and acceptance.

Mediating strategy #3: Regulating emotions

Closely related to test practitioners' expectation management is the strategy of regulating test takers' emotional processes.

At the second The Extraordinary Leader workshop the participants' emotional state was a central theme and something to be shared with the others. As the consultant, Michael, said: 'I am interested in where you are emotionally', followed by a poll about 'what are you currently feeling?' (Figure 14).

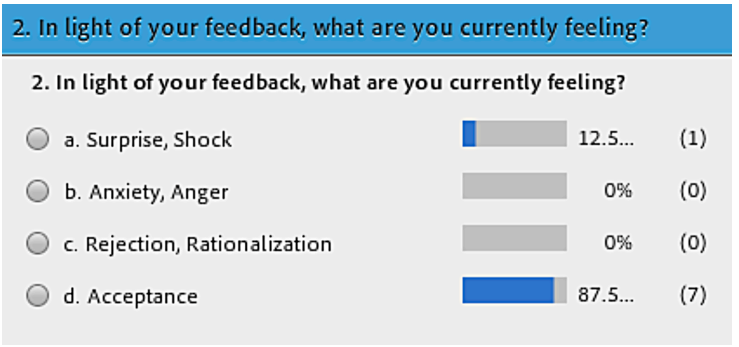


Figure 14: Poll: 'In light of your feedback, what are you currently feeling?', virtual workshop

The poll was used in both groups, but the results were similar: all but two, one in each group, chose 'd. Acceptance' as Figure 14 indicates, to which the consultant responds 'that's fantastic, that obviously makes it easier', and 'it is good with acceptance, then you can truly embrace the results'. Using a poll with few options, where participants are forced to share their emotional state, visually shows that acceptance is the 'correct' answer to the question. The poll establishes in a very

unambiguous way, that the majority of participants find themselves (or claim to be), in the desirable acceptance.

Shifting perspective to that of test takers' show how they adopt the emotional guidelines offered by Michael. Some test takers even talk about their emotional process using the same terminology as the one Michael used at the workshop. In the interviews, there are signs of a managed emotional process, and acceptance as the inevitable end goal. When I ask Tim if he remembers his first reaction to his 360° report, he responds:

> Yes, as a matter of fact I can. Then I could recognise this pattern of emotions they had presented to me. That you will deny until you accept and all that. And since we went through that you can say: "Oh yes, that one, that can't be right" and stuff like that. "That is not correct". But I got over that pretty fast and moved on to the acceptance phase, compared to if I hadn't been introduced to how we were going to react. We are not that intelligent after all, even though we think we are the most intelligent on earth. So, there is a pattern. And it is extremely great that we get that [introduction] beforehand, and say "everyone who receives this will go through this to some extent". (Tim)

According to Tim, the presented emotional pattern helps him. He evaluates, reflects, and regulates his own process by the help of the SARA stages.

John also describes his process by referring to the SARA stages:

> I went through the entire emotion spectrum. Absolutely. I was, when I read it the first time, in shock. Because I didn't really stand out on any of them, and I was lying relatively low on these different 16 or how many there were, what are they called, areas that you get assessed on. So, I was completely hit by a shock phase at the beginning and then later a reaction: "this can't be right, someone must have answered incorrectly". You definitely go through these things. And then I thought: "fair enough, we were told this". Make sure to look through it and learn, and we have to be careful not to say: "it is just because someone didn't understand it or answered something else than we wanted them to". (John)

Both Tim and John express how they self-regulate their emotional process based on the presented SARA phases.

How the participants handle their feedback appears to be regulated through different mechanisms, such as disclosing them on a set scale in a plenary, and by managing their expectations in the first workshop session. Presenting what they

ought to experience and feel serves to guide these expectations, which then makes them unwrap the feedback 'gift' in the desired way.

For The Extraordinary Leader, another site of emotional regulation is the one-on-one coaching session between the consultant, Michael, and the test taker, Joseph. For example, Michael read and translated Joseph's results in a deliberately positive way. Michael continuously praised Joseph, characterising some of his results as 'extremely impressive'. He was careful to maintain a positive atmosphere, for example by softening the statement: 'one item where you scored the lowest', by quickly adding 'you didn't score low on any of them'. Further, through adjustments, the feedback appears more positive than it perhaps is. Michael sorts the information and disregards certain inconvenient results. For one item with a rating of '1' (the lowest possible rating), Michael states:

> We just have to remember that when people receive this questionnaire, some say "well, I have just ten minutes, I'll rush right through it", and other times you might have just had a meeting with some colleague and he suggested something that you didn't think was a particularly good idea, so when he gets to this questionnaire then a quick rating at one, and... Sometimes it can be a bit random. That is also why we need to be careful not to over interpret the individual items. (Michael)

Later, Michael disregards ratings from the 'other' group, justifying this by describing the group as a 'special category', which is why he would not 'read too much into that'.

In the part of the test concerning the commitment of the employees whom Joseph managed, one disclosed that he/she is actively looking for other jobs, to which Michael remarks:

> Either this person had slippery fingers and pushed the wrong button by accident, or someone is actually looking for other jobs, which there can be several good reasons for. Like "now I have been here for ten years and I feel like making a change" or "my wife has gotten a job in Jutland, so we are considering moving", or "I am ambitious and would like to advance in my career, which is not possible here at the moment". The bottom line is, there are many possible reasons that have nothing to do with you or PharmExtra. (Michael)

Michael rationalises Joseph's results, deciding and adjusting what is of significance and what is not. He is careful not to place the responsibility for other people's negative ratings, or low employee commitment, on Joseph or PharmExtra. In contrast, Michael makes strong efforts to maintain a positive, optimistic

atmosphere, and a coherent narrative by introducing hypothetical explanations for such negative ratings. Michael also sometimes downplays divergences between Joseph's rating of himself and his manager's rating of him. He goes as far as to suggest that Joseph and his manager might actually agree, even though numerically, it doesn't seem like it. Discrepancies, potential sources of conflict, or ambiguities that do not fit the narrative are, to a certain extent, disregarded as unimportant or unreliable.

Michael is also cautious when talking about Joseph's areas of improvement: 'If we take your Overall Leadership Effectiveness Score and say it's at 3.7 now, let's just play, let's say we want it up to 3.9 in twelve months'. Making the suggestion for development playful takes some of the seriousness out of it. Also, Joseph's future improvements are articulated in numerical terms, making the suggestion remain abstract, technical, and impersonal.

The importance of consultants interpreting the test results is also acknowledged by James, who is responsible for the leadership development programme at BigBank. He explains that his choice to arrange for test takers to get their Hogan Leadership Forecast feedback session with a consultant before they receive their test report is based on his experience where test takers were 'deeply shocked' about their results. He elaborates:

> People become paralysed, especially young, inexperienced leaders. They don't read behind. [The test result] is taken for what it says, and this is where you have to ensure that there is buy-in by establishing that it is something positive that comes out of it. (James)

According to James, the feedback session gives the test takers 'sort of a way out', 'even though the facts are still there'. In contrast 'the mechanically generated report [doesn't] give them that option'. James describes a need for softening strategies and emotional regulation, in order for test takers to psychologically respond to the measures without rejecting their results.

Mediating strategy #4: Silencing critical questioning

Related to test practitioners' efforts to establish the trustworthiness and authority of tests, they also tend to brush off or only half-heartedly give attention to critical questions. This tendency is prominent in both the interview material and observations.

At an interview, the First Vice President, James, at BigBank, explains:

> We would very much like to get him [the consultant] to tell us more about these competencies and Hogan, so we can demystify it, because it is engineers we are working with, even though they are leaders, and they really want: "What is the mathematical formula, how has it been calculated and decided that I am *there"*. And I just feel like does that really matter? We are using a tool, so just believe in the data. And then, back to how it feels like to be examined and measured and weighed, I mean, people are always, when things are good, then they are happy, and then the survey was good. When the results are bad, then it wasn't a good survey, and "I don't understand this" and "that is not me" and denial and projection and all these typical psychological things that occur. Yes, this is you. Just get on with it. (James)

James advocates explicitly for a form of blind faith, by encouraging test takers to accept the tool and stop asking technical questions. There are traces of annoyance in the above quotation, where James appears frustrated with critical attitudes towards tests and technical questions. This frustration suggests that, to James, the technical aspects are not really important. Instead, it is about making an active choice to believe in the measure. He argues: 'It is not a question of beating people over the head with a stick, it is a question of making them believe in it'.

James further explains why the test takers need to just believe: 'It [the test] is not supposed to be inhibitory. It is supposed to develop people. If you have all the defence mechanisms ready, then… Then there is a long way ahead to build or restore faith in the fact that this is what works'. According to James, an overly critical mindset obstructs the development of faith in the tool. Notably, it is faith and belief that the tool not only works but is *what* works that is essential in James' work to reduce critical questions.

Other times, test practitioners provide vague answers or diversionary responses to critical questions, silencing or confusing the sceptic. At the certification workshop at People Test Systems, one participant enquires about the possible conclusions one can infer from self-report tests:

> Participant: 'But isn't it just the person's own self-evaluation? We don't actually know if people actually talk a lot, how it is perceived by other people?'
>
> Karen (instructor): 'It measures the degree of self-awareness.'
>
> Participant: 'Again, isn't it just the person's own assessment? We don't know how it will be perceived?'

> Karen (instructor), (hesitating and then responding): 'We-e-elll. Then we will take a look at the parameters of trustworthiness.'

The instructor answers the participant's question by referring to the 'parameters of trustworthiness', which according to the educational material include: 'realism of replies', 'realistic self-assessment', and 'response coherence.' These three as a whole are taken to indicate how trustworthy a person's test answers are. If this score is below 60:

> …it is very likely that something is wrong with the result, and it is [the person responsible for the test's] recommendation that the test taker retakes the test. The reason might be that the test taker has tried to manipulate the test, or it can be because the person in question simply is not capable of assessing him/herself realistically. Finally, it may for some come naturally to exaggerate a little. (People Test Systems, educational material)

This reasoning implies that a person's trustworthiness score determines whether or not one can make hypotheses about how others perceive this person. In other words, if a person has a high trustworthiness score, it is reasonable to assume that one can predict how others might perceive that person. Karen's answer is based on the tool's premise and restrictions, she does not engage in a discussion of what one can infer from self-assessment tests. But by referring to trustworthiness scores, which are deemed to be numerical proof of a test taker's ability to answer truthfully and realistically, the discussion is stopped and the participant is left with the option of just believing in the tool and its trustworthiness.

Mediating strategy #5: Disclaiming other tools

Another way test practitioners (and the companies they represent) seek to establish the tools' legitimacy and superiority is to argue that they are more valid, advanced, and scientific than other tools on the market. In doing this, test practitioners do not simply point out how the tools differ, but disclaim other tools. Disclaiming other tools both took place in informal conversations that I myself participated in or that I overheard during my observations, in interviews, test material, and on websites.

In informal conversations, it was not rare that I (over)heard someone talking about the limitations of other tests and how there are numerous problematic tools 'out there'. What were considered to be 'simple' typological tools such as DiSC or Insights were especially ridiculed.

Also on websites, more or less explicit comparisons to other tools are made. On Hogan Assessments' website it is stated that: 'No other company has measured job performance to the extent and depth that Hogan has' (Hogan Assessments, online, n.p.), placing this measurement tool above all others. How Hogan Assessments has reached this result is not made clear.

Likewise, People Test Systems' website indicates that their tools are 'state-of-the-art test tools. Tools that show a more nuanced and valid picture of the tested person' (People Test Systems, 2019). It is unclear what these tools are compared with. It is possible that People Test Systems means that their tools offer a more nuanced picture than if no tool at all was used, or alternatively in comparison with other tools on the market. Further supporting this claim, the website includes the information that '92% of their customers recommend People Test Systems', followed by selected reviews praising the company and their tools.

People Test Systems' claims on their website are further reinforced in their educational material. Here, they state in bullet points how People Test Person differs from other tests. This includes, for example, that it is specifically made for the business context, not based on American test systems (like other tools are), and that it compares job demands with personality. By highlighting these differences, People Test Systems indirectly points to the limitations of other tools.

Likewise attempting to elevate themselves above their competitors, Human Developers' educational material reveals that their incentive for developing their measure was their 'growing reservations concerning traditional tests currently used in industrial and organisational life'. Describing 'growing reservations' in connection with their choice of developing HD Leadership, suggests to the reader that Human Developers is an observant and critical company, and more importantly, that they have resolved or succeeded in working around the problems of 'traditional tests'.

In the participant manual for The Extraordinary Leader, Zenger Folkman argues confidently that they have 'moved beyond the traditional approaches to personal development that have been proven to produce "average" results' (p.2, module 1). Aaron, the Leadership Development Specialist at PharmExtra, who administers The Extraordinary Leader appears to share this opinion and states in an interview: 'Previously we've worked with tests where it's just yourself answering. And we didn't really feel that that was sufficient since it is then very subjective' (Aaron). According to Aaron, tools based on self-report are not sufficient due to their

subjective character, also suggesting that the 360° tool that PharmExtra uses is the opposite: sufficient and more objective.

The disclaiming of other tools was strongest in interviews with test practitioners, who defend the scientific legitimacy of their own tool while criticising that of others. Cathy, one of the psychologists who developed HD Leadership, explains that while other test companies are appealing, they fail to deliver high quality assessments. She argues: 'The problem is that they don't pass the test if you have to make a deep assessment and they can't... They don't go deep enough if people really have a problem and you really have to figure out what it's really about. Then they often miss it' (Cathy).

One of the main reasons why Cathy finds these other tools problematic is because they are not actual personality tests (as opposed to Human Developers' tests). She claims that they mask themselves as personality tests, but are actually measuring behaviour. In the interview, she remarks about other test developers: 'I can hear that they don't even know the difference when I talk to them at conferences', indicating that they are less knowledgeable than Cathy or even ignorant about the distinction between personality tests and behavioural assessments. According to Cathy, the latter is less precise:

> If you use the DiSC or Garuda or Papi or People Tools and all of them, that's behaviour plus competence. When you work with behaviour and competence, you are not as precise in the object you are measuring. (Cathy)

By arguing that other tests are less precise and their developers less competent, Cathy makes her case that HD Leadership should be the preferred tool.

Jacob, one of the consultants representing Hogan Leadership Forecast, argues instead that this is a better tool, especially compared to most others:

> Hogan is a pretty advanced tool in many ways, but there are assessments that are, well, sorry but they are just terrible, right. It's all about finding the truth. And then you suddenly happen to be blue, and then you just have no chances here in life … Like DiSC for example and PI, which is often used as a screening tool, it's very superficial. I think, even Myers Briggs, I think if you look into the validity of the test, you feel like crying. But you know, some people think it's fantastic … I don't want to bash anyone. But I think for some purposes, you might as well just use a horoscope. (Jacob)

According to Jacob, the validity of some tools is so low, that a horoscope might do the same trick. Paradoxically, Jacob points to a non-scientific instrument such as a horoscope as an alternative to claimed scientific test tools, suggesting that in some respects, the level of validity or science is not important. What's important is having an instrument that claims *something* about someone.

Miles, who also represents Hogan Assessments, criticises other tools by referring to their 'poor validity':

> I am very result oriented and therefore I also care a lot about what validity is and all that and that is also what guides my perspective on the individual tests. Because tests may well be visually attractive, there are test systems like Insights where you have some colours and you are weighted differently on the different colours. When I look at the validity, it is very poor … as soon as you start making major decisions on it, I would say, "slow down, now we've put people in boxes that are too big", so we have to believe that people are just a little more sophisticated than what a simple system like that can tell us. DiSC is another example of a test, where the validity is also really poor. (Miles)

Miles argues here that making tests visually appealing is a way of masking poor validity. He warns against categorising people in 'boxes that are too big', since this reduces complexity too much (in contrast to Hogan Leadership Forecast). He concludes: 'There are an incredible number of tests that are completely hopeless'. His criticism of other tools is combined with his conviction that tests are and should be inescapable since 'there is no other way'. Miles then quickly adds 'but you have to use the right tool'. Since there are many 'hopeless' tools on the market, Miles argues that one needs to be careful and pick the 'right' tool, in this case a tool developed by Hogan Assessments.

Representing yet a different opinion, Michael, the consultant for the 360° tool The Extraordinary Leader, argues that you need to stick with multi-rater tools, stating: 'There are five, seven, eight global [360° tools] which are really, really good, and it doesn't matter which one you choose. And then there are tools which are really terrible, and you need to stay away from those'. In contrast, Jacob argues that such tools are 'not very valid' since they are dependent on the situation. He explains that 360° tools give a snapshot of the person which relies on raters' moods and personal opinions. Jacob concludes: 'So a 360 is much more what-oriented. What happens here? Hogan has the potential, it is more concerned with why is this happening?' Thus, Jacob manages to endorse Hogan Leadership Forecast, while disclaiming 360° tools.

Despite test practitioners constructing arguments that they claim are built on data and reliable knowledge, some acknowledge that test choices rely on commitment and belief. Michael argues for the use of 360° tools by stating that 'self-perceptions are not as accurate as others' perception', which is in direct contrast to Jacob's opinion. In establishing Hogan Leadership Forecast as a precise tool, Jacob argues that 'the most valid is people's own self-perception'. However, Michael also acknowledges the significance of faith. At a lunch meeting he states: 'It is like religion; you believe in different things'.

Practitioners' test-subscriptions appear to be closely linked to what they consider 'reliable belief'. The belief is considered 'reliable' since test practitioners claim that their arguments are built on solid, trustworthy ground, especially compared to others' arguments. Test practitioners themselves, with the exception of at least Michael, might not acknowledge that the basis of their arguments may resemble religious belief. Instead, they refer to data supporting their arguments.

Whatever test practitioners believe, or acknowledge to believe, they demonstrate their own knowledge, critical mindset, and ability to evaluate other tests by comparing measures to each other, pointing out their differences and other tools' shortcomings.

Overall, test practitioners manage through (at least) the five presented strategies, how test takers' process, react, and work with their test results and thus create the foundation for acceptance.

Reflecting on the mediating strategies: Underneath the value-neutral surface

As shown in the section titled 'Normalising potentials', the tests consist of visual imagery and value-laden language expressing ideas of normalcy and desirable behaviour. In spite of, or exactly because of this, test practitioners attempt to frame the tests as value-neutral with harmless intentions, indicating that practitioners either believe the measures are indeed value-free with the purpose of initiating dialogue and reflection, or at least want them to appear this way. However, when digging a little deeper in conversations with test practitioners, value-laden purposes also emerge, illustrating discrepancies between test talk and test practice. The test practitioners' mediating strategies are thus informed both by test takers' various possible reactions and by the value-laden nature of the measures.

Despite Julie emphasising how the purpose of HD Leadership is not to instrumentally utilise the appraisals, later in the interview she mentions that Logistica uses an ideal profile that helps them evaluate test takers' results. This ideal profile, or 'target spans' is a mechanism Human Developers can enable, making one's chart look like this:

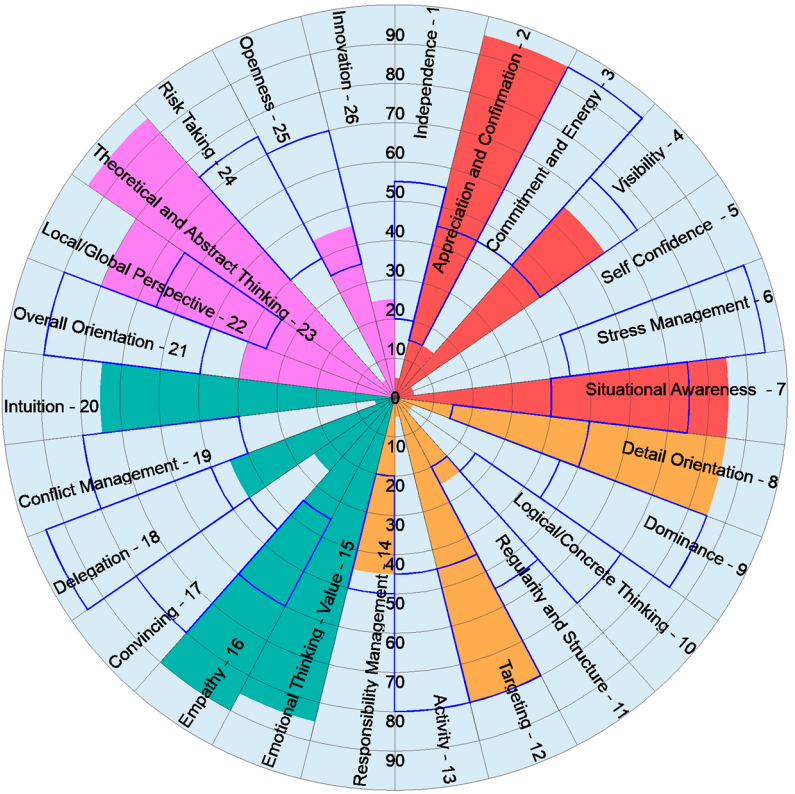


Figure 15: 'Overall summary graph' with leadership target spans, own HD Leadership report

The inserted target spans visually show how far my scores are from the spans that are deemed suitable and desirable for a managerial role. Indeed, the lack of colour in the target span fields creates very visible voids and illustrates my incapacity to achieve the ideal scores. As Human Developers states, the ideal profiles can 'specify the requirements for the leadership role and the development process the

person is facing' (Handbook for Human Developers Test System, p.63), reflecting a somewhat instrumental, evaluative purpose.

The test developer, Cathy, mentions this ideal profile in the interview, where she explains how it is based on:

> …theory and what you can read a leader should be high on … and then it is also about testing it in reality, to go out and test the stars ["great leaders"]. What do the stars look like? The leaders who are successful, what do they look like? They are also in the ideal norm or in the target span. (Cathy)

Interestingly, the leadership norm is built on 'what you can read a leader should be high on', which is exactly what the test developers are creating themselves: material telling others what a leader should score highly on. Moreover, a fixed leadership norm can only be created if test developers believe there are leadership 'stars', whose observable behaviours or personality traits can serve as inspiration, independent of context. The ideal profile is therefore, in accordance with test material in general, stripped of context, based on selections, and carrying traces of subjectivity. The ideal profile is based on 'theory' and so-called leadership 'stars', suggesting that the development of such a profile depends on what theory is deemed relevant, and who one believes to be leadership stars. These two elements are likely to be influenced by one's personal experiences and convictions, and dominant leadership discourse, such as transformational and charismatic leadership. Regardless, test developers are careful to explain how their theoretically based convictions about good leadership are 'tested' in reality. This is yet another way of instilling trust in their judgement and conclusions: Human Developers' knowledge base and ideal profile are tested and therefore sound.

Nevertheless, to compare Logistica's managers to this ideal profile, contradicts Julie's statement that there are no good or bad test profiles. The ideal profile, capturing the personality or behaviour of the 'stars', expresses an ideal to strive for. When such an ideal exists, profiles can, and should, according to Human Developers, be evaluated in terms of how well they fit this ideal.

Instead of using target spans based on theory and 'stars', companies can also choose to create their own ideal profile. Cathy explains:

> We will go in and say "what does your product manager look like?" Then they go in and look at the last ones, what, 10,20,30, or the top ten and bottom ten, and then see. Then you make the ideal profile based on that… They made one for [Logistica's talent programme], what they [the test takers' profiles] should look

> like, and then we go in and help them if they want it, and otherwise they [Logistica] go in and make them [the ideal profiles] themselves. (Cathy)

The process of identifying the people in the 'top ten' and the people at the 'bottom', is described as a straightforward and unproblematic activity. One simply identifies the best and the poorest. Besides exemplifying the normative nature of leadership, and the normalising potentials in leadership measures, the statement clashes with tests' value-free appearance as created and emphasised by test practitioners.

Julie further elaborates on how the ideal profiles work in Logistica. She states that the ideal leadership norm reveals what span on dominance a manager should ideally score within. This is not shared with the test takers. Julie explains: 'It is not a result as such, it is not like "well you score 35, so you are home safe". You can still be a terrible leader or a really good leader, regardless of that'. Even though it is not a result 'as such', Julie and her team still consider very high or low scores on certain parameters as potential challenges, since the test takers' scores might then 'go against certain characteristics that can make it difficult for them to be in a leadership role' (Julie).

Another way the test is used in Logistica is to construct team profiles: Test takers' results are pooled together, from which tendencies are identified. Julie points out that test takers are anonymised in the team profile: 'They are just dots on a scale' (Julie). For the purported sake of anonymity, individuals are transformed to representational 'dots', small parts of a bigger pattern. In this exercise, individual nuances are removed in the name of comparison and accumulation.

As mentioned earlier, the language in People Test Systems' educational material is value-laden and prescriptive. When I ask about this in the interview with Elizabeth and Emma, Emma responds:

> I think that's probably one of the things we would like to change, the written material, because it might be a little bit rigid, but there is also... there is also a balance to be found in terms of how we get someone to remember the interpretations if we don't describe them a bit simplistically. So, it also has to do with learning. We somehow have to spell things out for people to be able to separate things and understand it. (Emma)

Following Emma's rationale, People Test Person's educational material has to be communicated in simplistic ways. Emma considers nuances and gradations to be too confusing and difficult for people to remember. For the purported sake of

comprehensibility, a complex phenomenon such as personality is simplified. Regardless of this rationale, the educational material functions as practitioners' first encounter with the test and the meaning of different scores. This is what these practitioners bring with them in their future work with People Test Person. The reduction of complexity, in the name of practicality and user-friendliness, will nevertheless influence future test takers' understandings of both themselves and others.

A similar scenario is evident with Hogan Leadership Forecast at BigBank. There are discrepancies between the value-neutral front stage talk and how tests are spoken about behind the scenes. Megan from BigBank emphasises at the community meeting that the purpose of using Hogan Leadership Forecast is to create a common frame of reference and support personal development. However, in spite of this officially stated purpose, people at BigBank are measured on their ability to master eight chosen competencies: 'driving strategy', 'developing people', 'integrity', 'inspiring others', 'time management', 'positive attitude', 'team work', and 'accountability'. Through people's test-responses, their level of, for example, 'integrity' is evaluated on a scale from 1-100. As the First Vice President, James, says in an interview: 'We are trying to say, this is where *we* as an organisation want to go. How do you fit into that?' Evaluating how someone fits into an imagined ideal organisation is clearly a normative endeavour.

Later James elaborates on the eight competencies: 'It's a personal thing here. It's not something you should be forced to share with others because some will have a bad experience and get a wake-up call and say: "Fuck! I have something to work with here"' (James). Despite stressing the voluntary nature of participating in the competency assessment, James argues why it is highly encouraged: 'It would be stupid not to do it, and I keep saying it in these communities: "Go get these competencies assessed. It will give you an indication of where you are"' (James).

Even though the eight competencies are communicated as helpful for the individual's development, managers are measured quantitatively, on a scale from 1-100, and the results have consequences. In the interview, James recalled a manager with a rating that was 'very, very good on seven of the parameters'. However, since the last competency 'integrity' was low, James could not conclude that this manager was doing well. According to James, a low integrity score might suggest that this manager is good at exhibiting some qualities, but without any integrity.

This raises two points. The fact that James describes high scores as 'very, very good' indicates that high scores on those eight competencies are indeed desirable,

the norm to aim for. Also, a low score on integrity, a parameter perceived as important, can invalidate or undermine the scores on the other seven competencies. This evaluation of one parameter over others imbues it with significant powerful properties. The fact that some scores are considered better than others, might not come as a surprise. What is interesting is the contradiction between the front stage talk and the ways tests are spoken about behind the scenes.

We see the same contradiction at PharmExtra with The Extraordinary Leader. During lunch, Aaron explains how the 360° tool is not an actual test, since there is 'not a right and a wrong'. In contrast, Michael, the consultant, states in the interview with me that 'I have had a 360 report in front of me [and I was] thinking "this person should probably look for another job"'.

Aaron here represents the front stage talk about measures' use and purpose: that they are not actual tests, with right or wrong verdicts leading to serious consequences. However, Michael reveals the contrary, namely, that test results can indeed lead to consequential evaluations of someone's suitability for a position. Michael's statement mirrors Zenger Folkman's arguments that huge differences exist between top performers and average performers in any job category' and that 'the top person performing high-complexity jobs is 127 percent more productive than the mean average person, and infinitely more productive than the 100th person in that curve' (Zenger & Folkman, 2017, p.2). Differentiating between top performers and average performers indicate that there indeed are standardised, normative ways of identifying desirable and undesirable scores and test results.

Again, of interest here are the contradictions in how test practitioners discuss and present the test. They deny or downplay normative aims and try to disassociate the measure with a traditional test (where there are right and wrong answers), while also disclosing how the measures can reveal problematic behaviours and help identify mis-hires or the like.

Representing an empirical exception, Miles, a representative of Hogan Assessments, talks about ideal leadership profiles in a very blunt manner: 'Today if people say, [adopts a falsetto] "well, there's no good or bad personality" you can almost hear how I'm saying it, right, "for a leadership role". Yeah right' (Miles). Miles' imitation of other consultants' value-neutral front stage talk suggests that he finds them naïve or that they are deluding themselves. He elaborates:

> I could place brackets and say "this is the optimal area that the person has to score in", and then knowing that you will never get the perfect one. But let's say that, for example, as a leader, you have to be relatively high on power, not too high. You

> have to be outgoing, ambitious, maybe not too much, but if you are high on outgoing and ambitious, that is, you are competitive and driven by the specific tasks, project, functional responsibilities, then you can be a little lower on power. (Miles)

In contrast to the other test practitioners, Miles speaks about ideal leadership in quite unambiguous terms – apart from the vagueness of the advice to be 'high' in power, but 'not too high' and outgoing and ambitious, but 'not too much'. According to Miles, he would be able to identify the exact ranges for parameters within which a leader should score. Interestingly, on Hogan Assessments' website, the test development company Miles is representing, they state: 'There is no such thing as an ideal score or personality profile' (Hogan Assessments, online, n.p.).

There are indeed discrepancies between how tests are talked about and how they are used. This partly informs test practitioners' efforts to mediate the measures and regulate test takers' expectations and emotions. Moreover, the discrepancies between test talk and test practice are significant in understanding test takers' responses.

Responses to normalising potentials and mediating strategies

Test practitioners' efforts to mediate users' test experience are motivated by the variety of responses test takers can have to the measures and their normalising potentials. By instilling trust in the tools, regulating expectations and emotions, silencing criticism, and disclaiming other tools, test practitioners work to ensure that test takers accept the tools and fail to notice any discrepancies between test talk and test use. The variety and unpredictability of test takers' responses are key in understanding the rationale behind test practitioners' mediating strategies. The focus will therefore now shift to that of test takers – outlining the responses to both the normalising potentials and test practitioners' mediating strategies.

I first met test takers, who would later be my respondents, at the virtual workshop at PharmExtra. Prior to arriving I was unaware that the meeting was virtual. This upset my previous plans for how I would I introduce myself and how to approach the participants in breaks. I would not be able to talk to the participants, let alone observe their reactions to the introduced material and the consultant's messages.

I could not even put a face to a name, these names simply appeared as a list on the computer.

The unavoidable distance there would have been when I physically met the respondents for the first time, I would have normally tried to moderate or reduce through eye contact and informal conversations, with the expectation that my respondents would then be more comfortable at a later interview. But this format didn't allow me to do that, resulting in additional distance between me and my future respondents. However, I also thought that this particular online setup was interesting to observe in itself: How do people interact and participate virtually? My attention became focused on how the consultant introduced and framed the measure, since there was not much else to observe.

My first impression of the test takers at PharmExtra therefore solely consisted of the colour of their engagement beam and their contribution to the public chat. I next met these test takers one on one, either at a physical interview in a PharmExtra meeting room, or online, usually with video. The online format had again its advantages and disadvantages. Especially when respondents did not use a headset, the sound quality varied from decent to horrible. Sound issues led to many awkward moments in some interviews when I had to ask respondents to repeat their answer, or simply misheard them, leading to even more awkward moments. I therefore quickly added a request to wear a headset in the email invitation sent to respondents.

Besides the poor sound quality at times, I could again not make eye contact. Although we could see each other (more or less clearly), there were limits to how deep a connection we could make. With that said, generally the same social codex-following behaviour applied for these online meetings, as for physical ones: The level and type of chemistry varied, there was usually some small talk before the interview began, and we would exchange encouraging nods and smiles, meaning that the interview often felt relaxed and natural.

One of the biggest advantages of the online format was its flexibility. There was no need to book meeting rooms and no travel time. Meetings were easily set up since they could take place whenever and wherever, which was very important to most respondents. This flexibility was imperative with test takers working abroad.

I interviewed some respondents from PharmExtra at physical meetings. I was usually met with a very calm and composed person, in quite formal attire, which made me very aware of my own casual look. As if the respondents' clothes were indeed a costume, the interviews felt at times like a struggle to get behind the

façade. The distance almost felt bigger in the physical meetings, than at the online ones.

As a medicinal company, PharmExtra's main building at their headquarters appeared fairly sterile and modernist: clean, white, and uniform. Inside this building there was a lot of light, white-painted walls, and no ornaments or other knickknacks. The clinical feel was in stark contrast to my own office at Lund University, stressing the fact that I had entered strange land far away from home, perhaps with different rules and norms that I risked not knowing how to follow. In contrast, at the online meetings, both my respondents and I were placed in our own home court, so to speak, where any potential differences in the surroundings were invisible.

The test takers at Logistica, I only met online. Very few of the respondents turned their camera on during the interviews, obviously causing some additional distance. However, the biggest difference between the interviews I did with respondents from PharmExtra and those with people from Logistica, was the significance of my own interview experience. When I started interviewing PharmExtra-respondents in Spring 2018, I had done very few interviews alone. I was also still at the beginning of my PhD, meaning that my focus was still very much developing. When I interviewed respondents from Logistica in the Autumn 2019, I had done around 30 interviews with both test takers and test practitioners. My focus was sharper, and I felt more confident and comfortable as an interviewer. However, in both situations, the subject of test use and test experience seemed to be generally relevant and important to most respondents. This prompted them to open up and speak rather freely from strong positions (regardless of how I formulated my questions). On several occasions, test takers would move from short, diplomatic responses to suddenly sharing their opinions and theories eagerly and enthusiastically.

In general, the test takers were either somewhat sceptical of my research perspective on test use or found it refreshing and relevant. Those who were sceptical repeatedly asked for hypotheses, solutions, and psychometric assessments. To my best ability, I tried to explain how my focus was interpretative, and not evaluative per se, while taking note: Many test takers expressed a strong orientation towards instrumentalism and a need for clear-cut, unambiguous answers. Other interviewees appeared reassured and almost hopeful that a study like mine was taking place. These test takers generally shared substantially about their test experience(s).

Respondents frequently expressed strong feelings concerning test use. Some even became emotional. One 360° test taker, interviewed over the phone, told me that he was certain his manager wanted to replace him. He said he felt sick to his stomach after taking the test, and that he would often look in the mirror and wonder if he was good at anything at all. His manager generally gave him lower scores than his other raters, which worried him. The interview changed from one with preformulated questions to an informal conversation between two people, where we talked about a situation which affected him deeply. Suddenly the questions I had prepared were not that important. What was important was to just listen, make him feel heard, and maintain a safe space for him to share his thoughts. Ironically, the recording of this interview failed due to technical issues. As frustrating as it was in the moment, in a way, it also felt appropriate. This material was so personal and emotional, allowing me to refer more loosely to it here, but escaping the scrutiny of coding and analysis.

I interviewed all test takers after they had been tested, meaning that either staff from their company or external consultants had introduced and framed the instrument to them. More specifically, they had been exposed to consultants' mediating strategies. I observed that some test takers adopted the terminology presented by the consultant and/or test, buying into the premise of the test, regulating their emotions, and adjusting their expectations accordingly. In turn, some test takers were critical of the whole process. They had reflected on their results and the consultant's messages and strategies, leading to some strong opinions about what constituted ethical and unethical test use.

Based on interviews with test takers I present here a typology of reactions and responses to tests, their normalising potentials and test practitioners' mediating strategies. I have arranged these into two main types: 'appreciation' and 'scepticism and suspicion'. This is not to say that test takers' responses are unequivocally either in favour of or critical of test use. Test takers' reactions to being tested are filled with nuances and ambiguities. The same test taker can be conflicted, appreciating parts of the testing experience while disliking others, which is why some test takers are represented in both sections. The typology illustrates how test takers' responses tend to move across a spectrum from appreciation, joy and excitement about tests, to more critical, sceptical views on quantitative assessment tools.

The purpose of exploring the range of test takers' responses is to discover how they react in ways anticipated by test practitioners and in less predicted ways. Test practitioners attempt to be one step ahead of the unpredictable nature of test

reactions, the many nuances, and the occasional contradictions in responses. Practitioners try to foresee and manage test takers' reactions, causing them to mediate the measurement activity as meticulously as they do.

Appreciation

As the previous section, 'Mediating strategies', shows, the test practitioners' role – approach, style, and framing activities – is significant in establishing the tools' legitimacy, trustworthiness, and acceptability. These strategies are used to make test takers buy into the measures and accept and comply with their behavioural prescriptions. The strategies appeal to test takers' underlying assumptions and preconceptions about the value of quantitative assessment tools.

The quantitative test format itself thus attracts test takers, with some even arguing that the tools are objective truth-tellers (Connor; Sebastian; Layla), or a way of doing a 'sanity check' (Leo). Layla reasons: 'I think unless someone's telling you verbally on a daily basis or weekly or monthly basis, you kind of need a tool to tell you objectively to some degree "this is who you are"', ascribing a substantial mandate to quantitative assessment tools, due to its perceived objective character.

By referring to the tools' foundation in data and their numerical format, some test takers argue that the measures (and the HR departments administering them) become more credible than discussions and subjective assessments (Connor; Carl; Sean; Samuel). As Sean states: 'It's a great way to visualise. How are you really doing with this competence, sizing it up. How else would you do it? That would be very difficult. To sit and talk about it, I think'. According to Sean, a test is not just appreciated, but necessary. Without a test, assessing how one is doing 'with [a] competence' would be difficult. One would then have to resort to 'talking' about competency development, indicating that Sean finds this approach inadequate or not as trustworthy as when a test quantifies and visualises one's competencies.

Many test takers explicitly express that they consider data, scales, and measurements necessary components when assessing and improving performance. For some, this is because 'we love facts today, we love measurements, we love numbers' (Sean). This conviction is strengthened by placing quantitative assessments in contrast to bias, subjective hunches, and opinions. Rachel explains: 'We see that in PharmExtra, we always see that we have to be data oriented. Everything should be, every judgement should be judged on facts and data, not on feelings', leading her to the conclusion: 'Everything is really data oriented, so

it's no surprise that for assessing something, you need to have some figures, some data' (Rachel). Further supporting the need for quantitative data, critical questioning is referred to as 'suspicion'. Leo explains how he cannot use tests in his own team, since his employees would become 'suspicious' about how and why the test would be used.

Because they trust quantitative assessments, several test takers found measures helpful in justifying decisions and opinions, and confirming their progress and improvement. Using 'facts and data' is in direct contrast to 'making things up', as Connor, a HD Leadership test taker, argues:

> The test adds weight and allows you to say "yes but it's true he has to develop this dominance because he is low compared to his peers" and that's probably what it does. It gives it some credibility. So it's not just based on me making something up about you not being dominant enough. Then you have that one to refer to and say "it is true". (Connor)

As Connor expresses in this quote, tests legitimise decisions and opinions: By referring to test results, the need to justify, argue, and convince is reduced. Numbers allow test users to make claims which, without this numerical basis, would probably seem offensive. If someone is told that they are not dominant enough or that their mental resilience is too weak, without any referral to an external source of (quantitative) information, this could conceivably lead to significant conflict. Through the use of numbers, such claims are simply 'true'. Connor's quote also referred to a situation where a test result confirmed his impression of a person. The test was seen to add credibility and establish this opinion as 'true'. One might wonder what would happen in a situation where the test does not correspond with Connor's personal opinion.

Connor generally emphasises the positive side of measures' instrumentality and their establishment of statistical norms. According to him, HD Leadership is useful and effective when a test taker recognises 'a particular behaviour that they would like to change and then actually changes it for the better' (Connor). This is exemplified by his colleague who, according to Connor, would benefit from 'developing more dominance'. Overall, Connor finds HD Leadership (and its normalising potentials) to be a helpful guide in determining someone's leadership effectiveness:

> If you are going to be an effective leader, then the idea is that if you are very far from where the majority of other leaders are… you might be in the wrong place.

> So there are some things where it, you could say, where I think it is very clear which direction one needs to go in, but it has that… Well, it does not show it that explicitly, I mean, it does it in a slightly hidden way by showing that wheel. I mean, it's not like high is good and low is bad because sometimes low is good and, you know, it does this very well. (Connor)

According to Connor's own statement, the test guides how he evaluates himself and his colleagues by placing test takers in relation to a norm. Connor appears to believe that residing outside the norm means that you are 'in the wrong place'. Interestingly, Connor describes the prescriptive component as working in a 'hidden way'. The hidden or, at times, denied presence of normativity in the measures is a pronounced theme in my empirical material. As demonstrated in the section titled 'Mediating strategies', test practitioners spend significant time and energy framing their tools in a value-neutral manner, stressing how no scores are better than others, and that the tools are not meant to sort or categorise people as such. However, as Connor observes, there are hidden or implicit prescriptive standards in the tools; some more hidden than others.

Several other test takers share Connor's impression that tests are more valid than opinions and observations. Harry says: 'I think they [tests] have the tendency to be accurate and more neutral than basically human... yeah, well, human observation' (Harry). Samuel explains this in more depth:

> [The test] allows a conversation to take place about certain areas of people's personality. It's sometimes a difficult conversation to approach. Sometimes people don't recognise it in themselves so it's impossible to have that conversation and sometimes you can talk to someone and you haven't really convinced them of your opinion, sort of thing, whereas this objectivises that conversation. (Samuel)

According to Samuel, test results serve as convincing documentation that one can point to in order to objectivise a claim or opinion. Without it, some conversations are 'impossible' to have.

Erin has the same impression:

> I do believe at least modern managers or modern leaders believe that this gives a certain level of independence, you can say. It reduces subjectivity and personal bias and personal opinion in terms of "this manager is good because I like that person". It takes a bit of that away, right, and it gives some... (Erin)

In very explicit terms, Erin explains here why she thinks quantitative assessment tools reduce subjectivity, bias, and personal opinions. She doesn't finish the sentence on what tests offer instead, but I imagine that she is thinking of the opposite: that tests remove subjectivity, and offer a more impartial, credible, and reliable touch.

Leo expresses a similar opinion:

> As soon as you don't have something, some SMART targets, then it becomes a little more fluffy and just something you talk about, but you cannot really substantiate it. Why is one better than the other? Well, that's because "I think so". It's such a weak argument, whereas "he's a 3.7 and the other is a 3.2", well, then he's better. (Leo)

Leo attaches here a lot of persuasive force and objectivity to quantitative assessment, contrasting them with 'weak arguments'. Without measures, he appears to believe that all attempts of persuasion are lost causes. On the other hand, with measures, one can conclude with certainty who is 'better' at something, even if the difference is a mere 0.5.

Respondents appear to believe that the tools transform personal opinions or arguments into objective, substantiated claims and that credibility and objectivity are built-in features of the tools. Tests are perceived by many test takers to represent a contrast to (negative) statements and opinions that might otherwise be met with frustration or hurt feelings. Many of the test takers openly state that, by referring to test results or using test terminology, such statements and opinions become less personal and subjective and, as a result, more objective and impartial.

Besides using test results as 'weight', in other words, devices serving to legitimise decisions and opinions, some test takers also describe tests as providing 'proof' of improvement and change. In response to why Sean thinks quantitative assessments have become so popular, he says:

> If we can make something completely concrete and tangible, then we can show progressions afterwards. "See. You started here, you weren't very good. Now you have reached this point. That is fantastic". The fact that you can show this visually, right. After all, there is something about that format itself. We can measure development. We can give people grades in school and say "you started low, now you are there". It's probably some kind of human need. A visual proof if you can call it that. Visibility at least, of progress. I think that's what this is all about. And that's probably what everyone agreed on. "Oh, it's probably hard. How can one

> measure it? Human values, soft stuff, and oh, it's actually difficult. Nevertheless, how can we quantify that?". (Sean)

According to Sean, leadership measures answer the call for visualisation and illustrate proof of progress. Interestingly, he emphasises the value of quantification while at the same time contemplating the difficulty of quantifying phenomena such as leadership. Sean doesn't try to resolve this apparent problem. Instead, he acknowledges its difficulty, but somehow considers the challenge overcome. He has chosen to believe in the tools, despite his awareness of problematic or difficult aspects. The appreciation of facts, measurements, and numbers outweighs possible concerns.

Other test takers likewise emphasise the tools' ability to track and prove change and progress. Carl explains how the 360° report helps him because 'you actually get a rating, all the time [you] measure yourself and evaluate if you develop yourself as a leader' (Carl). By saying 'if' one develops and improves, he suggests that the evaluation depends on a measure showing a (numerical) change. Nathan also describes how HD Leadership has helped him assess whether or not there has '*actually* been a development', whether he has '*actually* succeeded with [his] efforts' (Nathan, emphasis added). Similarly, Catherine sees HD Leadership as proof: 'Everybody in those kind of positions wants to prove that they're performing or prove that they're improving and so a number makes that very easy'. These statements imply that development and success cannot be identified, communicated, or perhaps most importantly, trusted, without visual, numerical 'proof'.

Even though HD Leadership is a self-assessment tool, allowing for a subsequent test response strategy in accordance with the test taker's development plan, some of the test takers still consider it possible to objectively track a change (Nathan; Catherine; Connor). Along the same lines, Harry argues that leadership measures are used because they are sources of 'confirmation', both personally and as a political asset: 'I think an important aspect which we should not underestimate is the political aspect: Being able to demonstrate change is a really strong thing these days' (Harry).

These statements reflect the view that leadership measures are strong personal and political assets, based on the rationale that the tools help convey that one has improved one's performance by lifting one's score on a specific scale (Nathan). This is seen in opposition to feedback statements that are 'less measurable', making it more difficult to 'conclude [whether] it has improved' (Dan). Sean

likewise distinguishes between 'systematised' (quantitative) and 'nonsystematised' (qualitative) feedback. Quantitative, systematised feedback is thus perceived by some test takers to be more concrete, reliable, and tangible as opposed to qualitative forms of feedback.

Besides working as proof of change and improvement, the measures are, for some test takers, also knowledge sources of 'what's normal'. Thomas, after receiving feedback on his 360° report from Michael, reports how the session helped him infer 'this is normal, this is not normal'. More specifically, 360° test takers speak of changing their behaviour due to their test results, after realising that some types of behaviour are more desirable than other types. Rachel says: 'Thanks to the 360, I realise that I can look closed to people… And so I realise that I need to still to look open and maybe just to say "Okay, I'm sorry but can we reschedule something, we can see your point later," always be open' (Rachel). Based on both high and low scores, Rachel would 'try to compare and see what [she] could change'.

In general, the quantitative format of the tools and test practitioners' emphasis on numbers contribute to test takers trusting that their test results provide an objective and fair foundation for decisions and progress assessments. This is based on the conviction that the tools are valid, reliable, and objective, and, to a certain extent, that they tell the truth. Several test takers, primarily those who undertook The Extraordinary Leader, mention nervousness about 'the verdict' (Martha) of the test and how to deal with and 'swallow' the 'truths' (Sebastian) that can be 'hard to escape from' (Freddie). Vera perceives her test results as something that can 'prove or disprove' what her strong competencies are. The reason behind the attraction of quantitative assessments is here explained by Harry:

> There is a tremendous challenge in believing in qualitative data. Whenever I bring something qualitative up, including observations… It is easy to challenge qualitative analysis. People won't believe that it is a recurrent problem then they pay fortunes to get McKinsey to say the exact same thing. (Harry)

According to Harry's testament, the fact that quantitative assessments come off as believable and hard to contest is strengthened by the perceived shortcomings of qualitative data, which in turn is 'easy to challenge'. Qualitative data needs further legitimisation in the form of 'experts' validating, or simply repeating it.

The test practitioners' mediating strategies, applying numbers and highlighting their objectivity, thus fit conveniently with the test takers' pre-existing

appreciation of numbers. However, there was also scepticism and suspicion amongst test takers, which tells us that mediating strategies do not always succeed.

Scepticism and suspicion

As shown in the above, many test takers find that the measures and the test practitioners' messages confirm what they already presume about the need for quantitative assessment tools and their value. However, some of these and other test takers also respond to the tests' prescriptions and test practitioners' mediating strategies with scepticism and suspicion, often showing awareness of what is underneath the value-neutral surface. These test takers express an uncertainty about the trustworthiness of the measures' construction and conclusions, and a suspicion of the role and purpose of the test and the data it generates. For some test takers, their scepticism leads them to acts of (mental) distancing or more active counterstrategies, such as gaming the test.

On the surface, test takers appear to adopt test practitioners' front stage talk about the goals of the tools. It is some test takers' impression that the use of these assessment tools is meant to create dialogue (Daniel), inspiration (Eric), reflection (Oliver; Ethan; Daniel; Layla; Samuel), insight (Richard; Carl), awareness (Oliver; Sean; Layla; Catherine), a shared language (Charlie), to identify gaps and areas of development (Richard; Nathan; Joseph; Harry), improve collaboration (Nathan), open conversations (Samuel), systematise and structure feedback (Sean; Dan; Layla), and essentially, to contribute to personal development. However, in the interviews, a number of test takers also recognise and critically reflect on the behavioural prescriptions and values imposed by the tests, that is, the tests' normalising potentials.

Test takers speak about the tools being used to compare and sort (Nathan; Harry; Noah), assess suitability for attending talent programmes (Layla; Noah), and assess profiles' correspondence with 'ideal profiles' (Harry). For instance, even though the official purpose of Hogan Leadership Forecast is to provide a frame of reference and support individual development, one test taker claims that 'they've obviously been looking for a certain type of person' (Charlie). Charlie goes on to compare test results with giving management a 'loaded gun', which they in turn 'might not know how to use'. Similarly, Leo says:

> There are some leaders who pull it [the test results] out once in a while and it may well be that they don't pull it out at all because I've just been totally disqualified

> due to some delusion that I don't realise myself. Then it may well be that I just see it as a development area, but that it actually means that I will never advance here, and that aspect is quite unknown. You do not know what it is actually used for and it's obviously a little bit uncomfortable. (Leo)

The unknown aspect of test result use and interpretation worries Leo, making him wonder if he is just 'paranoid'. According to Leo, test results have an immense power, by potentially being able to disqualify people and prevent promotions. Along the same lines, Harry is concerned that, after receiving certain test results, he is written off as 'an emotional cripple'. He explains:

> Now I'm comparable. Now I'm comparable with all managers, right, and whatever I did in the last five years doesn't count because I am one of the people with the lowest emotional intelligence quotient, and the number stands. The number stands, no matter which context, right. The number stands. (Harry)

An effect of these measures is that the numbers are inescapable, which for some is experienced as a threat, or a 'danger', as Charlie, another test taker, put it.

Samuel explains the unavoidable reality of numbers:

> It's on a piece of paper. I can't deny it. I can either lock it in a drawer and pretend it doesn't exist or maybe I have to reflect on it, so it's not... You know what we're like. Someone says something and it bugs you or you just dismiss it and you don't hear it. Whereas a piece of paper, something in front of you, you have to reflect on it or you have to engage in conversation about it. (Samuel)

Following Samuel's reasoning, conclusions drawn from assessment measures cannot be denied. The format and presentation of test results forces one to engage with it.

Another concern for some test takers was the risk of creating self-fulfilling prophecies. Harry argues for example, that test profiles risk 'stick[ing] in his [manager's] mind'. Harry concludes: 'Okay, then we are making decisions that basically can harm my personal development based on my profile of one year ago. Oh, fantastic' (Harry). In Harry's experience, test results can lead to the very opposite of positive development and improvement. Rather than pushing people forward, tests can, in his view, hinder personal development. By establishing or repeating patterns of, for example, how and to who tasks are delegated, test results become self-fulfilling prophecies and maintain the status quo. Interestingly, when I ask Harry how he tries to avoid this status quo maintaining behaviour himself,

he admits to arranging his teams according to their test profiles. There appears to be a strong impulse to 'stick' with impressions and presumptions offered by the test results – even for those who voice a critical awareness of the problematic aspects of test use.

Another concern for some test takers in Logistica was whether or not ideal profiles exist. Around half of the HD Leadership test takers either had an unconfirmed feeling or said that they knew for a fact that Logistica uses ideal profiles in order to evaluate individuals' test results, whereas the other half were convinced that these do not exist. In response to my question on why Daniel thinks Logistica had chosen to use HD Leadership, he responds:

> If you speak in [a] pure [talent programme] context, then I have no doubt. It is my personal belief that they are looking for certain types of personality traits and these are the ones they would like to promote internally within the company. And of course, you can find this by using a tool like this one. (Daniel)

He adds a little later:

> They want to breed certain types of talents, you know, a certain personality type of talent, because there are also other talents who are really good at something, but who did not advance in this programme. I think that is because they were looking for some specific personality traits. (Daniel)

Daniel concludes that he assumes the tools are used to identify 'Logistica DNA', which is 'not communicated publicly of course', speaking to the hidden desirable profiles in measurement activities within the process or organisational leadership development. Similarly, Nathan has the impression that Logistica has:

> … some kind of template that says "well leaders should be in this span". That's my notion of it, it's unconfirmed rumours. So, that there is… You have to lie from, you know, from 45 to 85 on this scale, then it's good and if you are just below, then that's good enough. (Nathan)

Albeit based on 'unconfirmed rumours', Nathan expresses a strong feeling that Logistica has a more or less formalised conception of how a leader ideally should score on HD Leadership.

For Leo, he instinctively evaluates what parts of his profile are good or bad: 'When I look at the outcome, you know the spider web ['overall summary graph'] there, for example, I clearly have an impression of what would be... where it is I should

move, what would be better' (Leo). Leo's orientation towards good and bad scores suggests that the test's normalising potentials (in combination with the mediating strategies) have provided him with a clear impression of how he needs to change. When I ask him what his assessment is based on, he replies: 'Yes, that's a good question. If it is just a norm, I mean, it's not an explicitly stated criterion from management or anything, that one must be something particular' (Leo). Leo cannot specifically identify why or how he has come to the conclusion that his 'spider web' should look different, indicating that the normalising potentials and mediating strategies have worked in subtle ways to make him understand that he needs to change his behaviour, while he maintains a belief that there are no explicit demands (possibly generating less resistance).

As opposed to feelings, hunches, or guesses, Harry expresses more certainty about the existence of ideal profiles:

> Well, we had been told that there are certain traits which are good for leadership, right? So, in some areas, if you would score low on them, then you... the suggestion was created that you… should work on that because those are traits that leaders need. (Harry)

According to Harry, desirable 'traits' or profiles are far from hidden, rather, they were communicated explicitly to him; an experience the other test takers did not share. Harry then continues to talk more generally about the 'indoctrination' of values in 'big companies', drawing critical attention away from Logistica:

> What happens in these trainings is an indoctrination of particular traits that the company is in favour with [sic] and ah, and look, "you have seen your test results and there you deviate. Hmm, that's interesting, maybe you should change," right? … So, the corporate indoctrination of what are the values we like to see here, right? There's always the underlying disclaimer of "you might want to consider to change". (Harry)

Harry doesn't explicitly link value indoctrination to Logistica. Neither does he think tests' 'pseudo-academic, pseudo-psychological' (Harry) characteristics apply to HD Leadership. However, Harry is the test taker who on another occasion talks about being reduced to an 'emotional cripple'. He thus appears to be quite critical of test use in general, but reluctant to criticise Logistica or test use too much, suggesting that test practitioners' attempts to silence criticism has influenced Harry, further reinforcing his reluctance to raise concerns. Harry reflects on his hesitancy to voice criticism and concludes that he will have to pick his battles:

'Well, I would have to sink a very powerful boat and believe me, I will choose my fights wisely, I won't do that' (Harry). Instead, Harry places responsibility on himself: 'I will simply accept that I have done too many tests, I will always embrace test results as they are. I don't take this one too serious [sic] and I will basically try to be a good citizen in the company' (Harry). Being a 'good citizen' is important to Harry and implies that he does not air his criticism openly:

> I will speak more open [sic] if this stays within the two of us but if I know it's going to be shared… I won't speak more open, right, because I want to be a good citizen of my company that I work in. But then, that doesn't mean that I can't be critical with things. But I wouldn't necessarily say it everywhere. (Harry)

Interestingly, Harry equates being a 'good citizen' with obedience, compliance, and acceptance, which are not typically associated with contemporary forms of leadership that emphasise outspokenness, out-of-the-box thinking, and creativity. However, a 'good citizen' does correspond with the SARA response phases and reaction management, where test practitioners encourage test takers to 'just believe', accept their results, and be grateful for their feedback gift.

Noah, another test taker, doesn't refer to an explicit ideal leadership profile, but he is still quite certain of its existence. When asked if he thinks there are particular desirable profiles or characteristics, he replies:

> I think it's safe to say that that is the case, yes, but the... the people who were using this tool to make an assessment about who can get on the programme and who can't would have looked at certain aspects and said "well actually this is a problem". So, I mean it's no... it's no surprise that people who exhibit extrovert behaviours tend to succeed more and yeah, these are... So yeah, I'm absolutely certain. I'm absolutely certain that they will have used this data in a way that, from a purely professional point [of view], they shouldn't have … Do I like that those results would've been used by people who were not necessarily professionally qualified to interpret them to make decisions around who got on the programme and who didn't? Yeah, I don't like that and I'm pretty sure that happened. (Noah)

According to Noah, test data should not be used to evaluate someone's suitability for a leadership development programme or a talent programme. He is certain that this has been the case though. Moreover, Noah thinks it is 'no surprise' that people with extrovert profiles succeed more, suggesting that he indeed sees beyond the value-neutral surface and acknowledges that some profiles are just more desirable than others. However, he rejects the idea that such information should form the basis of selecting participants for a talent programme. Convinced that

Logistica uses ideal profiles in this way, Noah finds inappropriate, he also doubts the competencies of those handling and interpreting the measures. Overall, this causes him to be rather critical of the entire measurement experience.

Despite the actual presence of ideal profiles and other more subtle leadership norms, most test takers deny their existence, perhaps due to their taken for granted character. Gabriel, who is sure there are no ideal profiles in Logistica, later says that he himself has a classically good leadership profile, since he is an 'extrovert', suggesting that, more or less consciously, he considers some characteristics to be more suited for leaders than others. However, Gabriel does not recognise this understanding as representing an ideal profile, indicating that he has experienced such strong socialisation that he doesn't notice these value-laden ideas and taken for granted notions about good leadership.

The same sort of contradiction or tension appears when test takers are asked about their relationship to the statistical norm. Gabriel, Harry, Joseph, Dan, Oliver, and Martha all argue that they, to some extent, have not been concerned with how they have been placed in relation to a norm. Some even state that the norm did not matter to them at all. However, they then usually begin to speak about either high or low scores. These terms are only possible *because* of their relation to a norm. A score cannot be high or low if it is not compared to an average. For example, Gabriel says, he has not been bothered with averages or norms, but then continues to explain how it can be important to measure and assess where someone scores on for example dominance, in comparison with others: 'Let's say that many managers score relatively high on dominance, if you then score low on dominance, do you then have a challenge, in, say, leading people with strong opinions?' (Gabriel). The statistical norm is perhaps associated with comparison, competition, and hierarchy, something the test takers might want to disassociate themselves with. However, test takers talking about high and low scores, and what is normal and abnormal, suggests that the normalising potentials have indeed guided test takers' attention.

In spite of test practitioners' efforts to frame their instruments as trustworthy and accurate, a recurring concern amongst some test takers is the arbitrariness of filling out the tests. These test takers point out the possibility that people 'probably just ticked a box' (Richard), that results depend on raters' or test takers' 'mood' or 'mindset' on the day (Nathan; Catherine), on how much test experience one has (Harry; Noah), or how individuals interpret scales and items (Sebastian; Sean; Martha; Vera; Thomas). Sean ties this concern about interpretation directly to the limitations of numbers: 'Numbers alone are a bit hard to move on from, because

you have tried to categorise these different leadership qualities, but there are many ways to interpret them, I think' (Sean). Martha expresses the same concern, when she explains that results depend on how people interpret being 'result oriented'. Or as Vera simply put it: 'It depends on how you interpret the questions'. These statements indicate an overall scepticism towards the activity of quantifying leadership qualities because of the range of possible interpretations. Sean points more specifically to the interpretation of scales:

> I am also fully aware that some of these numbers, depending on what you put into them, you should have some reservations. But it's interesting enough to see... What you also need to remember, when giving such a test to people and asking them to answer from one to five, is that there is a difference in how people respond to such a thing. There are some who are very black and white who answer "yes", "no", "a five". And then others really try to use the entire scale and differentiate. (Sean)

According to Sean, how one responds to test items depends on 'what you put into [the numbers]' and how one interprets a scale from one to five. This reasoning shows that he does not fully buy into test practitioners' claim that the tools are objective. More or less explicitly, he questions here the scientific trustworthiness of the measures, by arguing for their reliance on and influence of human interpretation.

Frank likewise expresses scepticism of the test's ability to produce 'realistic' results, when he shares his impression of how respondents fill out the 360: 'They don't take too much time to think about the questions, they answer very quickly, la la la, just to finish … So maybe it could be not realistic sometimes' (Frank). However, on a different occasion, Frank told me that the 360° experience prompted him to attend therapy, which 'changed [his] life'. Even though Frank has his doubts about the tool's trustworthiness, he still tells a story where the test results caused him to change his life (in a revelatory and positive way). It therefore appears that some test takers need not necessarily perceive the tool as 'realistic' or objective in order for them to treat it as significant and valuable.

In contrast, Oliver's scepticism has led him to question the test's validity:

> You can say, these kinds of tools are presented in a way, where it is of course emphasised that it is not the truth, but you still work with it as if it is. And what do I think about that? I just don't think that it is particularly factual, if you ask me. It's opinions and [a] "matter of the day". (Oliver)

Oliver shows here an awareness of test practitioners' use of mediating strategies by drawing attention to the ways tests are framed and sold by test practitioners and test companies. Further, the fact that Oliver has experienced how tests are not presented as truth-tellers per se, but treated as if they are, indicates that he is very much aware of the discrepancies between measurement talk and measurement practice. Oliver concludes: 'Considering how big this folder is, and how much is presented as if it is true science, then I don't respect it that much. In that case, I think it is presented as being more scientific, you know, than what I think it actually is'. Oliver contests the test, so to speak, and questions the claim made by Zenger Folkman that the tool is truly scientific.

Tim likewise shows awareness of test practitioners' mediating strategies, framing efforts, and front stage talk, when he states 'a good story has been created around it [the measure]', which he assumes is necessary since 'we are talking about perceptions. We are talking about feelings'. According to Tim, stories told around the tools are necessary since the tests cannot stand on their own; the tools do not hold enough convincing power in themselves.

Other test takers point to the limitations and problematic implications of behavioural measurements. In her interview about HD Leadership, Layla says: 'Maybe that's the issue, that the test measures what you believe and not necessarily what you do at all times'. She goes on to argue for the potential benefits of 360°s, of getting other people's perception instead of measuring people's self-assessment. She alludes to the idea that self-assessment tests merely measure people's 'belief' and where they prefer to place their mark. Another limitation of tests is voiced by Harry who says: 'I have seen people that are incredibly creative, that you could not read that from their profile [sic]', suggesting that tests might miss important characteristics or behaviours.

Yet another test taker points to linguistic issues in HD Leadership, resulting in her answering the test in a particularly random way:

> I had no clue what I was answering when I answered those questions because they didn't make any sense to me whatsoever. So that's the point, you know? [I had] totally no clue what they were going on about in any of the questions and, like I said, that was very much the feedback from all the native English speakers, actually, that have done the test. (Erin)

Erin has completed the test with 'no clue' about the questions she was answering. Working in HR, Erin is both a test taker and a feedback-giver. She gained certification, which enabled her to give feedback to test takers in the company.

However, she finds the language and formulations so incomprehensible or meaningless, that she has no confidence in her own test results, or in giving feedback to others:

> I can't explain that when I'm giving feedback to people. So I would still, even now, a couple of months later, I would still not feel competent doing the feedback on this particular test, you know? (Erin)

Possibly as a result of this, when I ask her if HD Leadership has had an impact on her life, thoughts, behaviour, she replies: 'No. Not at all' (Erin).

For some test takers, their scepticism takes the form of measurement fatigue, discouraging these test takers from engaging with tests in the way practitioners encourage them to: openly and enthusiastically.

Harry, Noah, and Gabriel describe how their measurement fatigue now influence their approach and attitude to tests. Their extensive test experience means that their knowledge about themselves is 'saturated' (Noah), and they now master the measures. As a consequence, these test takers explain how they find it difficult to gain new and meaningful insights from the test results. According to Noah, the extent of self-awareness and hyper-reflection means that he 'can't really answer neutrally', because 'you have that in the back of your head probably', leading to him 'skewing the results by answering questions in a very exaggerated manner' (Noah). Harry describes a similar consequence:

> Well, for instance I know where my problem areas are, what I have to work on. Therefore, I give already answers that will suggest that is the case and then... Because normally they would always say "yeah, just answer intuitively on the questionnaire", right? "Just don't think about it too much". Yeah, but if you have done like literally ten times the questionnaire, how can you not think about it? And if you know already the areas where you potentially have to improve because you can feel it every day in every situation, because you have been through that reflective process so many times, then you would give the answer either that you don't want to give or that you want to give. And if you're honest, which I am, I wouldn't give the answer that I think, or I think, I'm not sure, if it is actually the problem because I have not had a psychologist sitting next to me. It's only the test which is always telling it, right? So it's literally like a self-fulfilling prophecy and if you're answering the questions not intuitively but with a really good knowledge or an assumption about what the problem is, then you determine the result. (Harry)

Bringing about a self-fulfilling prophecy, Harry answers items based on knowledge from previous test experiences, which then reinforces his test profile. For example, if several test results claim that one is very detail oriented, one might answer future test items as if this has already been established. According to Harry, this is unavoidable, despite instructions telling test takers to answer questionnaires 'intuitively', indicating that previous test knowledge replaces intuition, or offers a sort of intuition about which boxes to tick. Indeed, one could argue that Harry does answer the items intuitively. Intuition relates to knowledge or beliefs that are without analysis, reasoning, or deeper reflection. Harry answers the questionnaire without reflecting more deeply about his previous experiences and how they might skew his results. In this sense, he simply fills the boxes instinctively.

For Nathan, the risk of self-fulfilling prophecies was a big concern the second time he underwent testing via HD Leadership: 'When I took it the second time, then I probably knew what to expect. I think then I had a concern that I would be biased and try to game this test since I know what factors I have to work on' (Nathan). However, Nathan continues to explain how he decided to trust the validity of the tool since there are 'so many questions and control questions', which he believed would make a gaming strategy impossible. He goes on to argue for the importance of getting a result that is as 'objective and correct as possible', which is why he said he tried to answer truthfully, without keeping in mind the 'bias' he mentioned.

In contrast to this strategy, Harry and Noah find that their extensive test experience inevitably directs the way they fill out tests, and ultimately reduces their learning potential. Harry even describes that he has now figured the test out. He acknowledges that he knows too much about himself and the tests. Harry explains how he has completed many different tests, leading him to be:

> … a bit fed up with these kinds of tests. I keep getting more intelligent, do you know that? I have the feeling I have understood the principle of how these questionnaires are running … I know what they ask, I know what to answer because I know how I am. But it feels like the result is more extreme, as if I would have just answered what I wanted to hear. And the result is basically always the same and it's getting more extreme. (Harry)

The combination of Harry knowing himself well and knowing the principles of the tests enables him to predict the results and therefore answer the items in a more extreme way.

Charlie describes the same tendency. However, Charlie sees his experience as allowing him to tweak and game the tests. Since Charlie is convinced that BigBank is looking for a 'certain type of person', he answers the test strategically: 'I was one of two that met the perfect profile, so that's nice to hear. So I got quite "hey, I've gamed it, I've got this strategic thinking right"' (Charlie). He elaborates:

> If I was a 20 year old doing my first assessment test, I'd have definitely answered a lot of things differently. I think we experience and know what people are looking for. Then you can sort of tweak it a little bit, if that makes sense … So I think the more experienced you get, the wiser you get with these tests. (Charlie)

The same phenomenon is present for all three test takers: They express having such extensive test experience that they respond to test items in a steered or extreme way. Whether or not it is a deliberate gaming strategy or an unconscious act, they perceive their results as expressing something skewed or adjusted.

In sum, test takers' responses to being measured: appreciating the activity, being sceptical and suspicious or even counteracting the intentions underlying the tests and their practitioners are what motivate test practitioners' mediating strategies. Despite official claims (on websites for example), about tests' neutrality and ability to objectively evaluate test takers, test practitioners are very much aware of the range of potential reactions to such measures potentially initiate. For example, Michael presenting test takers with the common reaction pattern (SARA), James carefully arranging the order in which test takers receive their test report and get their feedback, are efforts that speak to an awareness, perhaps an alertness, of test takers' responses.

Despite test practitioners' awareness of the possible responses to the measures and their consequent use of mediating strategies, some test takers remain sceptical. However, it might not be despite of the mediating strategies, but partly also because of them. Test takers respond to the entire measurement activity: the framing of the test, taking the test, and getting feedback on the test. Some test takers appear to be sceptical precisely because of test practitioners' efforts to guide test takers' measurement experience.

Of greatest interest in terms of numbers' capabilities to perform effects is the relationship between the test's normalising potentials, test taker responses and test practitioners' mediating strategies. Before I go deeper into this relationship in my discussion chapter, I will now tell my own test story, illustrating the personal and emotional investment and inner process which tests and their practitioners call for and initiate.

Putting myself to the test

I was tested myself three times in the course of this study and received feedback twice. During interviews, two consultants offered to send me the test and give me feedback after, which I happily accepted. The two experiences where I also received feedback (Hogan Leadership Forecast and HD Profile), are the ones I will share here. I took Hogan Leadership Forecast first and HD Profile second.

I completed Hogan Leadership Forecast in May 2019 on a Saturday afternoon at home. I read the instructions thoroughly, as if it was an exam that I could answer incorrectly. I was nervous although Jacob had explicitly told me in the interview that there are no right or wrong answers, which was also stated in the instructions. Maybe I was nervous about having to answer hundreds of questions as 'honestly as possible'; my instinct is to always respond to categorical questions with 'that depends'. Being forced to choose an option that might not feel fitting sparked some nerves. In addition, my test anxiety might have had something to do with a fear of being analysed, exposed, and figured out somehow (or that people *think* they have figured me out), and that my profile would be perceived as weak or problematic. More than anything, I feared being misunderstood and reduced to something I would not be able to recognise (or perhaps wanted to recognise).

Answering the 450-600 items was a mixed experience. Some items required almost no reflection time, some were very difficult to answer, and I simply did not understand the meaning of others. When filling out the personality inventory, I had to answer by choosing either: 'strongly disagree', 'disagree', 'agree,' or 'strongly agree'. In the motives/values assessment, I had to answer either 'agree', 'disagree', or 'don't know', this last being described as having 'no opinion about the matter'.

I found many of the items odd or even impossible to answer without any context, especially for people with 'untraditional' jobs or people who are self-employed. For items about one's relationship with one's boss, everyday interactions with colleagues, and one's ability and desire to take charge in projects, I found it difficult to answer as a doctoral student. Some items that puzzled me were for example: 'Fear has been an important driving force in my job'. What was the meaning of this? That fear got me to where I am, or that it drives me to keep going? And fear of what? Another item was: 'Our Post Office is pretty inefficient' which I found difficult to answer, wondering what my answer would say about me. More examples are: 'In my opinion you can't trust a person who drinks' and

'My friends know how to party'. What would my answers to these items say about me?

The items that caused me the most trouble were those that, in my opinion, largely depended on the context, such as 'I would typically take charge in a group' or 'I like speaking in front of a lot of people'. Even though the instructions said to answer the items according to how you are 'the majority' of the time, I still found items such as these hard to answer. Being forced to answer with one of three or four options was frustrating: I felt the need to explain and provide additional information.

Hogan Leadership Forecast comprises three assessments. I finished them all and sent off my answers to the consultant. I looked forward to getting the results, but also wondered what Jacob would think of them. And would the results stick with me in a way I would rather avoid? The latter concern was probably related to the fact that the test includes a reputation dimension, where you are told how people perceive you, solely based on your own answers. This dimension added an extra layer of pressure: On the basis of my answers, the consultant would (claim to) know how other people perceive me. According to supporters of the test, including the consultant I interviewed, Hogan Leadership Forecast is based on a massive amount of data which has been thoroughly validated, so I felt that potential resistance or scepticism might be difficult to voice. Also, such supporters believe that one's personality is stable, but not static, meaning that my test results are assumed not to change unless I experience a life-altering incident. Was my argument for so many of the test items that 'it depends' then disqualified?

Approximately two weeks went by before the test results landed in my email inbox. I had planned to read systematically through every section of the results. However, the amount of information quickly prompted me to skim through parts and wait for the feedback session with Jacob to try and make sense of it. This decision was further supported by the language in the report. Some of the conclusions or hypotheses were quite categorical. At times it felt like I was being told, by some distant, nameless authority, who I was. Some 'developmental recommendations' included formulations like: 'When talking to direct reports, make sure to listen; talking is not always communicating' and 'don't be defensive about negative feedback,' which felt somewhat like a reprimand. I did not necessarily agree with all of the recommendations (or was that just me being defensive?).

At times, formulations went in different directions. For example, the report concluded that I both like working with others and working alone. While these

are not mutually exclusive, other results contradicted each other. For instance, that I like and dislike public speaking. These contradictory conclusions reflected a contextual component within the test. However, it also seemed like a type of safeguarding – the results could this way match anyone.

Two weeks after I received my test results, I had an online feedback session with Jacob. Trying not to come off as defensive, I awaited Jacob's analysis. The session took almost two hours and was different to what I had expected. The feedback became very personal and at times resembled a therapy session. However, Jacob was careful not to overstep any boundaries, and appeared genuinely interested in my interpretation of the results and his hypotheses. Jacob interpreted the report in what I felt was a meaningful way. He highlighted themes and conflicts from the report that I had not been able to spot myself. Interpretation was indeed key. I was left with a feeling of having been heard and understood – quite far from being categorised or labelled.

After the feedback session, I wondered how I might have experienced it if I had been rated on competencies, or evaluated with a specific leadership role in mind. The process and feedback would then have had a different aim. Since nothing was at stake for me, professionally at least, both Jacob and I discussed the results freely and with no 'ideal profile' in mind.

Having this experience to look back on, I was excited to try another test: HD Leadership. Cathy, the test developer, asked me in the interview with her if I wanted to try it, and she also offered to give me feedback.

I took the HD Profile test in November 2019 during a normal work day. Similarly to the earlier test experience, I felt a little tense when reading the instructions in the email:

> Please answer the questionnaire within the agreed time period. There is a total of 271 statements that must be answered in one stretch, as far as possible. Remain undisturbed while answering. Choose the answer that comes to mind first. You cannot correct a response you have submitted. It takes most people approx. 30 minutes to answer the questionnaire.

Already there was a statistical norm to consider. Answering 271 items in 30 minutes meant that I roughly had to go through nine items per minute, in order to be among the average test takers, that is. I wondered what it would say about me, if it took me longer to complete the test. The reminder that I could not correct a submitted response did not ease my nerves either.

My nervousness was quickly replaced with feelings of frustration and confusion. This test comprised so many items whose meaning I did not understand. These included: 'When I work on an issue, I usually consider the principal matters first' and 'I prefer to analyse my way to the principal matters of an issue'. What do 'the principal matters' cover? And what is this preference an alternative to? Or: 'Parts are more important than the whole for my understanding of things' and 'Abstract (theoretical) thinking is a prerequisite for practical decision making'. What kind of decision making? I would think it depends on what the decision is about. There were many of these types of formulations: 'When assessing an issue, you always have to consider the concrete first'. What is 'the concrete'? And: 'Life is full of opportunities and tasks to be solved'. How are opportunities related to tasks to be solved?

I hesitantly answered all 271 items and was then left with a feeling of uncertainty. Since I did not understand several items, I answered them quite randomly. I wondered whether this would show in the results. And if I would even be able to recognise anything about myself from the generated report.

Two months and two reminders to Cathy later, I received my test report. The front page displayed this greeting:

Figure 16: Front page, own HD Profile report

My test results are personal, but paradoxically enough I was reduced to a number, something notoriously impersonal. I was now known as '99337'. What I also found interesting about this introductory greeting is that test practitioners often emphasise how quantitative assessment tools are complex and capable of almost anything. Oddly, substituting the number with my name has either not been possible or merely not prioritised. However, a number on the front page rather than my name might be a strategic choice. The impersonal impression initiated

associations to something technical, automated, and serious. Not unlike medical tests and results, I was reduced to a quantity, suggesting that I was indeed measurable, and that whatever information followed this page was reliable.

I scrolled further through the digital report, immediately noticing the very high and very low scores. I did this despite having heard during my field work that this is not the most constructive approach. But it felt instinctual to pay attention to high and low scores, both since these visually stood out and since I figured they would bear greater meaning and implications.

Shortly after, I received feedback from Cathy via Skype. This feedback session was very different from the one with Jacob. This time, I did not get the impression that the consultant was interested in my views and perspectives. Cathy cut me off several times, quickly moving from one scale to the next, as though my test personality consisted of boxes to be ticked off. She frequently asked me if I recognised the descriptions of my scores, only to interrupt me after it seemed she considered that I had either validated or invalidated the claims. She appeared impatient and in a rush. She did not seem interested in explanations, elaborations, or reflections. Before we reached the last part of the feedback, called a 'cluster analysis', where I assumed some sort of narrative would be constructed, we ran out of time. We made plans to go through the cluster analysis the following week, but Cathy never showed. I wrote to her, but never heard back.

The two test experiences were complete contrasts to each other, drawing my attention to the arbitrariness and human influence that surround test use. The test experience appears to rely on many different factors: one's previous test experiences (reservations, fears, expectations), the consultant's approach, and the comprehensibility and interpretation of the test items.

In my case, the test experiences were also influenced by the fact that I was taking these tests while simultaneously studying them. I underwent these tests two and a half and three years into my PhD, which has inevitably influenced the experiences. At this time, I had completed most of my field work. I therefore had, theoretically based assumptions about the phenomenon and empirical material that pointed to certain findings. Specifically, my assumptions and empirical material sensitised me towards techniques employed by the consultants and possible value-laden or paradoxical test items and phrasings in my test reports.

The combination of my own test experiences and my analysis of observations, texts, and interviews has led me to reflect on wider questions in relation to the use of measures in leadership development. What is the relationship between the

measures' normalising potentials and test practitioners' mediating strategies? How can the performative effects of leadership measures be characterised? These are some of the questions that I explore in the following discussion.

Discussion: The performativity of numbers

Measures aim to shape test takers' identities and encourage them to change behaviour according to the established 'normal' within the measure. But, as I have shown in the previous chapter, this is not the whole story. I have found that the link between quantitative assessment tools and individuals' responses and changed behaviour is a mediated one. In the context of leadership measures, numbers do not speak for themselves or perform automatically. The use of measures in leadership development programmes is a meticulously managed process, one that relies on certain actions and reactions from the social actors involved. In particular, consultants and instructors frame the measures and facilitate the entire measurement activity for test takers or future test practitioners. From the tools' official presentation on websites, to the introduction of the tools to test takers, to test feedback sessions, test practitioners mobilise norms that frame the instruments and regulate users' expectations, responses, and potential resistance. Through these efforts, practitioners guide test takers towards the 'normal' test experience, that is, a test experience that achieves particular goals: test takers responding to the measures' items in an honest way, accepting their test results, refraining from asking too many critical questions, and the identification of appropriate developmental areas in line with the organisations' objectives.

In this chapter, I unpack the link between measures and their performative effects. Drawing on Austin (1962), I suggest that numerical utterances, like speech acts, require certain conditions for their performative intents to be realised, and propose ways to explore how these conditions are achieved or hindered by social actors.

In developing this argument, I first consider how, in existing critical literature on the effects of quantification, researchers generally tend to underestimate the importance of quantification's contextual factors. Underestimating or downplaying the significance of such conditions impedes a full picture of

quantitative tools' performative effects. Second, in accordance with Austin's (1962) views, I make the case that the performativity of numbers relies on context. More specifically, I argue that the effects of leadership measures are a potentiality, whose actualisation depends on social actors' efforts to establish certain circumstances around quantitative assessment tools. In the third section, I discuss the mediating work quantitative assessment tools rely on, furthering our knowledge of the performativity of numbers, i.e. what the numbers require in order to carry out illocutionary acts and have perlocutionary effects. In the context of leadership measures, numbers alone do not greatly affect their users or targets, which is why it is essential for social actors to intervene and mediate the process.

The performativity of numbers in critical literature on quantification

As we have seen in the theoretical section: 'Social measures and their normative implications', critical scholars tend to focus on the performative power of commensuration itself or quantitative tools themselves (Cohen, 1999; Espeland & Sauder, 2007; Hacking, 1990; Mau, 2019; Porter, 1995; Rose, 2008). These scholars are concerned with measures' ability to normalise, discipline, change, and guide human behaviour and value systems (e.g. Espeland & Sauder, 2007; Hacking, 1990; Mau, 2019; Porter, 1995; Rose, 2008). Further, critical scholars point to actual (normalising) effects of quantitative assessment tools: When we measure, we nudge people's behaviour to fit the numbers (Adolf & Stehr, 2018).

What these perspectives share is a focus on the power of quantification itself. As Espeland's first sentence states in the foreword to *The New Politics of Numbers*: 'Numbers do things' (ed. Mennicken & Salais, 2022, p.vii). For example by transforming 'how we understand our selves' (Mennicken & Espeland, 2019, p.224). Earlier, Espeland and Stevens (1998) write that commensuration 'encourages' (p.323) 'transform[s]' (p.328) and 'produces' (p.331), that commensuration has 'constitutive power' (p.331), and that commensuration can 'radically change the world by creating new social categories' (p.323). Here, commensuration is the powerful subject performing ('encouraging', 'transforming', 'producing', 'changing'). Likewise, Hacking (1990) and Rose (2008) argue that numbers and data about averages have great power in that they promote ideas about normality and as a direct result, identify abnormal behaviour. Hacking (2006) further makes the case for how different 'engines of discovery'

such as quantification, have the power to 'bring new kinds of people into being' (2006, p.6) through new classifications. There is an assumption in the critical literature on quantification that numbers have an agency of their own, that they have the ability to transform the world and bring new things into being.

This assumption partly exists for good reasons. As mentioned in the theoretical section: 'Quantification, objectivity, and normativity', quantitative measurements are closely linked to scientific ideals of objectivity, rationality, and accuracy. The power and authority of quantification is therefore coupled with the authority of science. In other words, because of our faith in science, which is often equated with quantitative measurements, we grant an authority to numbers, consequently ascribing them persuasive force and allowing them influence.

However, jumping to study the power and effects of quantification, scholars in this field tend to ascribe a performativity to quantification in a way that suggests that numbers do their work by themselves. One might get the impression that quantitative tools 'change', 'transform', 'normalise', 'produce', seemingly without relying on outside assistance. The conclusion – namely, that quantitative measures can have powerful social impacts (Mau, 2019), forward agendas and guide behaviour (Cohen, 1999; Espeland & Sauder, 2007; Porter, 1995) – is entirely justified, but an important part of the puzzle is missing when the power of quantification is contemplated without full consideration of the measures' mediators and their strategies.

In the critical literature on the effects of quantification, the conditions quantification requires to have a performative effect are thus either downplayed or only briefly acknowledged. Scholars who recognise the conditions refer to numbers' reliance on interpretation, ascribed rationality, and the need for buy-in and acceptance (e.g. Espeland & Stevens, 1998; Gould, 1996; Mau, 2019). Mau (2019) argues for instance that numbers require a leap of faith. In other words, for rankings to be taken seriously, we have to be convinced of their value. As Mau (2019) further posits, the establishment of quantification's legitimacy and objectivity relies on the authority of the experts who present them. Likewise, Gould (1996) argues that numbers 'suggest, constrain, and refute' (p.106), but in order to do more, we need to interpret them – which can then lead to conclusions that favour certain (political) standpoints and prejudices (Gould, 1996). However, in these and other texts, numbers and quantification suddenly 'do' things again.

Thus, in literature on the effects of quantification, there appears to be a recurring assumption that the conditions needed for numbers to perform are more or less

automatically in place. The required contextual factors are rarely treated or empirically explored in-depth e.g. *how* faith in numbers is instilled, *how* experts express authority and establish their measures as legitimate – and the amount of time and energy this work calls for. Most attention in both critical and mainstream studies is allocated to the effects, powers and properties of metrics, rankings, and commensuration processes themselves.

As mentioned earlier, there are exceptions to studies that downplay the context quantitative activities rely on. One notable exception that my work builds on and extends, is Espeland and Stevens' paper from 2008. Espeland and Stevens argue, with inspiration from Austin's framework, how numbers' persuasive force is conditioned by the authority we grant them. The authors also emphasise that since numerical pictures are not, metaphorically speaking, transparent glass windows but images that 'color and refract what comes through' (p.425), they require someone who decodes the information and interprets the images. Espeland and Stevens therefore point to a number of contextual factors influencing the effects of numbers. My study extends these arguments by empirically showing the significance of contextual elements and social intervention. In contrast to Espeland and Stevens, my study draws attention to the mediating work quantification relies on in order to have performative effects, and less on what they refer to as the 'infrastructural' work quantification requires in order to be carried out (p.410).

Although developed in relation to linguistic analyses, Austin's framework resonates with my argument that quantitative 'utterances' rely on context to have performative effects. In other words, social measures, like speech acts, have performative intents whose actualisation requires certain circumstances to be in place. The conditions necessary for performative acts to be realised are key both in relation to linguistic utterances and quantitative utterances. Borrowing from Austin's (1962) work on speech acts, knowledge on quantitative tools' effects can thus be further developed.

Considering the felicity of numbers

As discussed in the theoretical chapter, successful or 'happy' perlocutionary acts – actually persuading someone, or actually making someone act – require certain conditions to be in place, such as the speaker having the necessary authority and being in the 'right' physical surroundings (Austin, 1962). Where Austin mainly

focuses on the conditions required from the outset for a performative intention to be realised, I find that, in order to actualise quantitative measures' normalising potentials, there is a reliance on continuous work, and that understanding this work is essential for appreciating such tools and their potential effects.

The significance of social actors' mediating work means that turning the tests' normalising potentials into actual effects is a potentiality. Indeed, what is typically described in existing literature as *effects* of quantitative tools are in fact *potentials* whose actualisation depends on social actors establishing certain conditions. Powerful effects of quantification are indeed not a given, as the literature often assumes. In other words, numbers do not automatically perform 'happy' illocutionary acts and thus realise themselves and their normalising potential based only on their appearance.

Suggesting that leadership measures do not just perform automatically invites us to consider them as performative potentialities. Drawing attention to the idea of potentiality means that the measures' effects can be achieved (in many different ways) or not achieved at all. Circling back to Austin, the performative outcome of an utterance is not a given. Austin describes how acts can produce both intended and unintended consequences, and that different effects can occur or not occur, regardless of the intentions of the speaker. He explains how 'we can import an indefinitely long stretch of what might also be called the "consequences" of our act' (p.106) meaning that there are numerous potential outcomes of an utterance and several ways on which the listener/receiver can react to the speaker's speech act. The field of possible consequences and the existence of an imminent 'misfire'-threat means that the illocutionary and perlocutionary performative acts are indeed potentialities. Uttering a number can have different, if any effects. Being told that there is a 30% chance of winning some game might persuade one person to play it, and discourage another, more risk-averse person. Also, the number must be uttered by the appropriate authority, in the appropriate context.

The concept of potentiality calls for explorations of the work social actors do to ensure the actualisation of numbers' normalising intents: the conditions required, and how these are established or hindered by social actors. Focusing our attention to the normalising potentiality of quantification means that the primary object of study shifts from quantification itself to contextual factors and the variety of responses from social actors. Acknowledging measures' normalising potentials means that we ought to look at the context in which they operate, to explore elements and dynamics that hinder, obstruct, facilitate, or support the measures' normalising aims. The concept of potentiality challenges the assumption of

quantitative tools' automatic and certain effects by pointing out that these effects are mere possibilities. More specifically, we need to look at the interplay between test practitioners' mediating strategies and test takers' (counter)strategies, since these determine how or if the normalising potentials of the measures can be manifested to perform actual effects. These ideas are unpacked in the following section, where I take a closer look at the conditions needed for the measures to have performative effects.

Creating the 'appropriate' circumstances

There are many factors that (can) interfere, obstruct, and hinder the normalising aims of the four measures. These factors pose a 'misfire'-threat, to stay within Austin's terminology, which is why test practitioners employ the five mediating strategies: creating legitimacy and trust, managing expectations, regulating emotions, silencing criticism, and disclaiming other tools. Moreover, in the context of leadership development, the appropriate circumstances involve how the measures are presented on websites or individually by test practitioners themselves, the authority and intentions of social actors representing and translating the measures/numbers, the physical or virtual environment where the measures are introduced and talked about, and how the audience responds to them.

Test practitioners work to ensure that the appropriate circumstances are in place before a number even appears. Through confident claims on the tests' websites and references to data, validity and reliability, test practitioners aim to mobilise the necessary authority and context needed for test takers to immediately believe the numbers when they appear and attach great importance and value to them.

The practitioners who are presenting the measures and conveying the tools' objectives work to ensure that test takers perceive the consultants as experts with the appropriate authority. When Michael (the 360-consultant), in flawless English, tells the participants that he has a background in psychology and finance and has lived in London for 12 years where he worked in investment banking, he seeks to establish himself as experienced, trustworthy, and dependable. Supporting this endeavour, Michael confesses to the participants that he himself had prejudices about the tool, 'but then [he] learned', making participants trust his experience and judgement and testifying that any scepticism will be refuted.

By referring to 'data', 'wise people', 'correlation theory and other mathematical stuff', 'strong and linear correlation', test practitioners make numbers appear technical and beyond manipulation, and thus 'shut down dissent' (Porter, 2019, p.595). This way, practitioners ensure that test takers ascribe trustworthiness to the measures.

The measures themselves are constructed in ways that radiate science and objectivity. In test material and on websites, the tests' validity and their various helpful purported abilities are repeatedly stated. Test developers frequently refer to the word 'data', appealing to our historically developed associations to rationality and objectivity (e.g. Cohen, 1999; Porter, 1995). Being quantitative in format, the measures appeal to our preconceptions about quantifications' superior status and fast track to certain knowledge (e.g. Espeland & Stevens, 1998; Mau, 2019). Test practitioners then carefully plan the introduction of a tool by creating attractive, yet serious-looking, physical or virtual environment, displaying convincing PowerPoint slides filled with graphs and references to 'data'. These efforts work to support the legitimacy and scientific status of the measures and the professionalism of the people behind them.

A lot of the test practitioners' work is put into managing the audience's responses to the measures, pre-empt resistance and critical questioning. As we have seen in the section called 'Responses to normalising potentials and mediating strategies', test takers react to the measures and the circumstances in a variety of ways. The prospect of test takers' opposing a measure and its prescriptions prompts test practitioners to frame quantitative assessment tools as reliable, trustworthy, and non-threatening, to create compelling and positive narratives based on test results, and to regulate test takers' emotional processes and reactions before they even occur. More specifically, practitioners establish norms about belief and trust, honesty and openness, gratitude, and acceptance, to nudge test takers' (emotional) process in the desirable direction. Test takers' reactions, attitudes and potential resistance are in this way taken into account in the efforts test practitioners make for the actualisation of the measures' performativity potential. Test practitioners frame and mediate the measures because of the many and unforeseeable ways that test takers might respond to these tools' potentialities.

Test practitioners' mediating strategies are thus not only supportive in terms of actualising the tests' normalising potentials, but also productive in that they work to create an impressionable, compliant, accepting test subject: the appropriate receiver of the quantitative utterance, the normalising potentials, so to speak. As Austin (1962) argues 'the procedure must be executed by all participants

completely' (p.36), by which he means that when saying for example 'I bet you sixpence,' the person to whom you propose the bet must accept it, in order for the utterance to be successful (p.36). Users of leadership measures must agree to play the game and accept the terms and conditions, which is why much of the test practitioners' work contributes to creating this compliance.

The performative potential within the measures invites a variety of responses to the tools. Test takers can choose to respond in ways that maintain the measures' normalising potential, actualise it, or transform it, encouraging all the mediating work of test practitioners.

These insights challenge quantitative studies on test use which depict test takers as either passive or reacting in ways that can be captured in a questionnaire with a restricted set of possibilities (Chan et al., 1998; Converse et al., 2008; Visser & Schaap, 2017). Indeed, test takers do not only (potentially) 'alter their behavior in reaction to being evaluated, observed, or measured' (Espeland & Sauder, 2007, p.11), they also alter their expectations to measures themselves, influencing how they receive and respond to quantitative tools from then on.

Test practitioners work in different ways to minimise test takers' potential resistance to the measures, that is, test practitioners stimulate trust in the tools and the numbers they generate. But, as we have seen, this is not all. Test practitioners also seek to prevent test takers from believing *too much* in the measures, by preventing or managing any excessive trust in the tools and numbers. What might seem like a paradox is in fact a strategy of persuasion, another way in which practitioners encourage test takers to accept the results of the measure. If test takers have too much faith in the tool, they risk ascribing it the ability to tell them an absolute truth, a truth that might be hard to accept. In other words, too much belief in the numbers' verdict can hinder test takers from accepting their test results. Test practitioners therefore also downplay the extent to which the measures are entirely objective, accurate, and reliable. What is interesting here is that test practitioners work hard to establish quantitative assessment tools as objective devices that can accurately capture one's leadership potentials and pitfalls only to later downplay or soften these abilities and any normative agendas of such tools.

When test practitioners deny the normalising aims of the test, it plays an important part of their work to moderate the blow of the test verdict and ensure that test takers will not reject their results and voice extensive criticism. Part of the measures' effectiveness is therefore denying the properties that critical scholars point to, namely the tests' normalising power. In doing this, test practitioners

stress the tools' soft qualities, moderating the tools' rational and factual conclusions. For instance, as we see in the section titled 'Mediating strategies', the 360-consultant Michael chooses what numbers to highlight, and what numbers to downplay or even discard. He explains away some numbers as the result of ratees' 'slippery fingers' or mood on the test day, suggesting that the tool cannot be blindly trusted, and that it is indeed not objective or completely reliable.

Test practitioners need to maintain test takers' buy-in and ensure that they stay committed to the measurement process, which is why numbers that are at risk of compromising this objective are mitigated. When practitioners soften some numbers' abilities to offer precise and objective diagnoses, or ignore numbers that are in the way of the overall performative purpose, these practitioners actually repress rather than actualise the numbers' normalising potentials. Repressing numbers' potential suggests that test practitioners do things with numbers that are not already allocated within the numbers themselves. The work test practitioners do around measurement activities is therefore not always to actualise their potential, but may include rationalising or discarding certain numbers.

The established norms and mediated trust test practitioners instil in their tools are essential for the numbers' normalising intentions to achieve the appropriate circumstances. In other words, numbers do not act on their own. When my scores in a test report gets me thinking and possibly make me want to change my behaviour, it is because test developers and practitioners have already contextualised and framed these numbers for me. The test practitioner has employed strategies to ensure that I trust the measurement tool, believe in it, and accept my results. Supporting these mediating strategies, the numbers are further contextualised in a report with a colourful, confident design and language, prompting me to take the numbers seriously and trust their validity.

More broadly speaking, numbers' performative intents do not actualise themselves. The number 14 on a piece of paper does not mean anything or do anything. Numbers need to be contextualised, framed, and expressed by a person, sign, or device in that way be *made* meaningful and not remain a 'lifeless figure' (Fauré, Cooren & Matte, 2019, p.354). When we check the temperature on a thermometer or on an app, the number is expressed by a trusted device. Finding out it is -15 degrees Celsius outside, we put on warm clothes or stay inside. This action is not initiated by the number itself, free from context or audience. The number '15' is first of all combined with a unit of temperature, Celsius, a known and established scale of measurement. Moreover, the number is conveyed by a physical thermometer, an app or a website that we recognise and trust. Lastly, the

context (season, people wearing warm clothes outside) corresponds with the number, making us, the audience, interpret the number as sensible and believable. Interestingly, quantitative tools, presented as objective, rely exactly on what objectivity historically has been conceived as being stripped of, namely 'inference, interpretation, [and] intelligence' (Daston & Galison, 2010, p.17).

In contrast to words, numbers do however appear to require less interpretation and to carry in themselves an authority of science, a persuasive power. As shown in the thesis' theoretical chapter, quantification is tightly associated with ideals of scientific objectivity and rationality. This association means that the conditions necessary for numbers to perform illocutionary acts are often readily in place. Quantitative expressions benefit from the scientific authority attributed to them. As we saw in the empirical section titled 'Mediating strategies,' when Michael, the 360-consultant, tells the participants that there is a 72% likelihood of someone becoming an extraordinary leader if this person combines the strengths 'builds relationships' and 'drives for results', he gets an immediate positive response to that number. The number '72' is not explained in detail. When a perceived expert simply utters the number 72 in the appropriate context, the number gains authority and makes the audience react.

Although at times they describe something personal, numbers are technical and less open to interpretation than words. If someone tells you that you are driving too fast, you might contest this and offer counterarguments. But if the speedometer indicates an excessive speed, you either slow down or actively ignore this number and break the law. The number on the speedometer does not function on its own, but it does invite less interpretation and fewer counterarguments. A number persuades us more easily than arguments based on words about our driving speed. Such words could be discarded as mere opinions, where a speedometer is typically perceived as expressing indisputable and precise facts.

The ascribed authority of numbers goes to the heart of the matter: Quantitative tools are in high demand precisely because of their numerical format. The instruments offer an alternative to qualitative approaches to leadership development, which many organisations regard as insufficient, inaccurate, and in need of numerical back up. Numbers thus do appear to have an inbuilt powerful potential which, under the right circumstances, can have performative effects.

With that said, the authority we grant social measures varies. In the case of leadership measures, their test developers tap into our association between quantitative measurements and the authority of science, providing them a

convincing starting point. As we see in the empirical chapter, our faith in science and preconceptions about the authority of quantification do incline many test takers to trust the tools.

However, this authority is not simply in place, or secured per se. Rather, social actors must intervene and mobilise strategies to continuously establish the authority of the measures. Objectivising the personal, the intimate, and emotional, the numbers invite different responses, where some are far from aligned with the otherwise strong faith in science-discourse.

Beyond the context of leadership development

It is important to note that some types of measures or quantitative utterances require more intervention to have a performative effect than others. The performativity of numbers depends on what the numbers are conveying and who or what is communicating them. Some quantitative tools and indicators are so well-established that they meet very little or no scepticism and critical questioning. These typically target physical elements, such as temperature, speed, distances, which we consider to be external. As I have argued, while these measures are contextualised in a way that grant them authority and meaning, their abilities to perform effects do not rely on substantial ongoing human intervention.

When measures then contain social or personal dimensions, more implicit political or economic interests, and consequences, they start to require more intervention. Overall, it appears that the higher the emotional stake and consequence of the measure, the greater the need for human intervention, mediation and continuous legitimisation of the tool. For example, both university rankings and leadership measures target performance and reputation. Neither are neutral, nor without consequences for institutions and/or individuals. Because of the personal investment one has in the ranking or test, these instruments are more likely to be met with emotional responses ranging from appreciation to scepticism (to flat out rage and uprise). To a larger extent, people contest the meaningfulness of ranking universities according to selected indicators or measuring someone's self-perceived level of empathy, since these are not fixed in any physical sense. Instead, they depend on value systems, human interpretation, experience, and assumptions. This is exemplified by Locke (2014) who describes, based on case studies, how senior managers at higher education institutions, in response to rankings, would work to 'maintain a degree of "stability" and agree a level-headed

and consistent attitude, "toning down" extreme reactions. They sought to "de-sensitise" the league table "issue" in the institution by "routinising" and "accepting them"' (p.85).

According to these findings, like leadership measures and personality tests, rankings require social actors to frame and mediate them, in the form of mitigation and reaction management.

Despite these similarities, notable differences exist between rankings and personality tests in terms of how much translation and intervention they need. Espeland and Sauder (2007) and Mau (2019) describe rankings as powerful, behaviour-altering mechanisms that do not just inform, but form. Mau, Espeland and Sauder all argue that information about rankings influences decisions, how people make sense of situations, and what and how administrators attach value. That rankings affect expectations and prompt people to change behaviour. Espeland and Sauder (2007) state: 'Rankings create self-fulfilling prophecies by encouraging schools to become more like what rankings measure, which reinforces the validity of the measure' (p.15). According to Espeland and Sauder (2007), these effects are closely tied to the power of commensuration (creating realities and guiding our attention). One possible explanation for these abilities lies in rankings' purpose and the scope of their potential impact. Rankings are supposed to determine the order of quality, in terms of fulfilling selected, measured criteria. The top university receives a score of 100, others then receive a percentage of this score, creating a very visible reputational hierarchy (Mau, 2019). Moreover, rankings are shared with a clear purpose of rising in the created hierarchy.

In contrast to rankings that target institutions, products, and services, scoring, the quantitative method used in leadership and personality measures, is an individual undertaking. The hierarchisation of value and quality and the prescriptive standards in scoring are more subtle and open to interpretation. The verdict of a leadership measure is less clear and absolute than a ranking of a university, which is why there is a higher need for framing and translation.

A test report does generate orders of worth, and it concludes what a person's strengths and weaknesses are, but the interpretation of scores remains much more complex than that of a ranking. The fact that people need to gain certification in order to give test feedback suggests that test results cannot stand completely on their own. Test practitioners need to interpret the tool's conclusions, allowing for different narratives to be developed, depending on the interpreter.

Moreover, leadership measures target the personal, emotional, and intimate. When something personal is described by something notoriously impersonal, numbers, my empirical material shows that tensions and ambiguous experiences are likely to occur. Social actors therefore have to establish these intimate numbers' authority and create their appropriate circumstances, repeatedly.

As a result, the performative effect of a personality test is, to a greater degree than a ranking, a potentiality. Commensuration, in the sphere of rankings, is perhaps more powerful on its own, given its immediate and visible hierarchisation of value. In contrast, the power of fixed numerical norms in leadership measures relies to a greater extent on social actors creating the appropriate circumstances and continuously linking the numbers to the authority of science.

In order to perform, some numbers thus require less intervention than others but all require some form of mediation.

In addition, different communication channels or people convey different numbers and measures. Many numbers that are shared with the public, such as Covid statistics, are ingrained in our institutions and communicated through official channels representing authorities. This is not to say that these numbers are never contested or questioned, indeed there are Covid critics or even deniers. Nevertheless, such numbers are surrounded by a well-established aura of scientific authority. A personality test is granted authority by its association with rationality and objectivity, but it is permissioned and interpreted by human beings. The person who permissions the test might be a colleague, while the consultant who interprets the measure may be directly facing the test user. The measures' advocates and those responsible for it are not abstract, distant entities, which ultimately makes the measures' aura of scientific authority less stable and as a result, more prone to being challenged.

Summing up

With the insights from this chapter in mind, I have extended Hacking (1999), Porter (1995), Rose (2008), and Espeland and Steven's (1998; 2008) arguments: Acts of quantification are social processes, tied to values and assumptions, and their performativity relies on social actors' strategies to mediate the measurement process and ultimately ensure the audience's acceptance and internalisation of the numbers' normalising aims. In relation to this, I argue that numbers have

powerful normalising *potentials,* which can be regarded as performative intents, rather than automatic performative effects. The measures' performative power depends on the broader social context and the norms and conditions established. The power and normative effects are not an inherent property of the measures. Instead, the tools contain a potential, that, if test practitioners frame and handle a certain way, can have normalising effects. The social actors are key. How consultants and feedback givers interpret the norms and numbers of the instruments and establish mediating norms around them determines how, what, and to what degree normalising effects take place.

My extension of critical scholars' arguments about the power of numbers and commensuration takes into account consideration of the conditions needed for this power to be actualised. This focuses attention away from the power in quantitative commensuration per se and instead draws attention to the (powerful) actors intervening or guiding the process. When scholars attach all capability or performative effects to the measures, as is often the case in studies where measures are claimed to either represent or create reality, the role of operators, mediators, facilitators, or interpreters, for example consultants and instructors, is overlooked or downplayed, whether intentionally or not.

In contrast, this study, where I have paid great attention to the types of mediating activities, focuses attention away from the power of metrics. After all, human beings develop the measures, sell them, frame them, and mediate them. Actors involved in these processes are key in understanding how quantitative social measurements work and normalise attitudes and behaviour. Moreover, through the concepts of *normalising potentials* and *mediating strategies*, I suggest *how* we can explore the field of quantitative assessment tools.

The concepts invite approaches and research questions that consider the performative potentialities, rather than given effects, of quantitative measures, and the work social actors do to actualise or reject these potentialities. When we explore the strategic efforts of social actors, the measures' reliance on and susceptibility to contextual factors and consequently the tools' various possible outcomes, emerge.

This perspective further allows for more critical questioning and acts of resistance, such as voicing concerns or suggesting alternative forms of personal development. If we better understand the conditions that measures require to have normalising effects, we can likewise better challenge, withstand, or counteract them. In contrast, when we assume and expect the measures to have substantial power in themselves, we attach a significant capability and mandate to them. When we

acknowledge that (some) measures' normalising, disciplining effects require considerable effort from social actors, they cease to be particularly powerful on their own, and signs of the tools' instability and vulnerability rise to the surface.

To sum up, this study contributes with extending insights to existing studies about the more hidden sides of social measures and their performativity: the (legitimisation) efforts on and behind the scenes (Elmes & Costello, 1992), the conditions required for numbers to perform (Dorn, 2019; Fauré, Cooren & Matte, 2019), e.g. the different actors involved in measurement processes, and their significance (Meier & Carroll, 2019). My study complements these studies by showing how measurement processes are indeed social dramas, where surrounding mechanisms such as soft norms and the actors mobilising them play a key role in how the measures perform (or do not perform) and are received by test takers. Like a theatre, actors give enthusiastic staged talks with a script-like feel, in order to convince and persuade the audience of their prop's (measure's) value. A successful show depends on the actors performing in a way that leaves the audience clapping, believing, and buying tickets for the next show.

Conclusion

Whenever I tell someone about my research topic, I usually receive one of two responses. Either 'that's interesting, I think there are way too many tests out there, and once I actually had a very unpleasant experience, where…', or 'that's interesting, so what kind of test should we use then?'.

The subject of measuring social phenomena tends to interest but divide us, and I have therefore had very different conversations with people in favour of and critical towards measurements.

These conversations, the research for this thesis, and the experiences of being tested myself have impacted my view of quantitative assessment tools. On one hand I am less intimidated by quantitative measures. I have become so aware of their shortcomings, odd rationales, and seductive means of persuasion, that I attach less authority to them. Instead, I perceive them, now more than ever, as rich sources of paradoxes, norm production, and examples of human beings' desire for categories, formulas, and truths. On the other hand, individuals in favour of and/or representing a particular tool have challenged my critical position with such conviction of the measures' value and necessity that I have felt the ground shake beneath me. These discussions have made me wonder about the legitimacy of and possible meaningful places for quantitative tools.

Regardless of both test developers' claims and my findings, some people find tests helpful, empowering, and even bordering on revelatory, while others experience tests as violating, misleading, or ultimately a waste of time. I myself have felt both the allure of tests and resentment towards them: the rush of adrenalin when I have received a test report and feedback, but also the feelings of frustration and resistance to the tests' instructive, so-called recommendations to change behaviour I could not recognise in myself, or the feedback giver's disregard for my own perspectives on my test results.

Fortunately, my quest in this thesis is not to indisputably determine whether or not we should measure phenomena such as leadership. Rather, I wish to share insights about and unveil the subjective, normative side of what is presented as an

objective, value-free phenomenon, so that people can relate to and make decisions about tests on a more nuanced basis.

Summary of findings

Through my study of four different measurement tools (HD Leadership, Hogan Leadership Inventory, The Extraordinary Leader, and People Test Person), I have found that leadership measures not only consist of and promote normative standards, but also rely on contextual norms to influence test takers. In other words, the measures are embedded in norms and normalisation efforts, efforts that are essential for the measures' normative agendas to be actualised. In the development of this argument, I have introduced the concepts *normalising potentials* and *mediating strategies.*

The measures have, as it were, in-built normalising potentials. Statistical norms decide if one's empathy level is high, low, or average compared to others. This information is then visualised on a scale, graph, or chart accompanied by possible recommendations to change behaviour. The four measures are in general loaded with value-laden language. Recommendations to change behaviour are not merely suggestive, but instructive through the use of authoritative imperatives (e.g. 'don't interrupt', Hogan Leadership Inventory). In spite of test practitioners' claims that the measures are not used to advance normative agendas, the measures themselves clearly promote and advocate for certain levels of characteristics, and thus encourage normalisation and conformity.

Beyond the encouragement of normalisation, any effects of the tools take the form of a potential that may or may not become actualised. Similarly to speech acts (Austin, 1962), for the performative intents in quantitative 'utterances' to be actualised, certain conditions must not just be in place from the outset, but must be continuously secured. Specifically, the measures rely on social actors' mediating strategies. These comprise the work test practitioners do to create legitimacy and trust in the tools, manage test takers' expectations and regulate their emotions, silence users' criticism, and disclaim other tools on the market. Practitioners frame the instruments in ways that appeal to test takers' preconceptions about quantitative instruments' superiority and objectivity, inclining test takers to expect valid, trustworthy verdicts from the test.

In line with the historical use of and perception of numbers and quantitative measurements, social actors who develop, sell, and use contemporary leadership/personality measures create auras of rationality, objectivity, and science around the instruments, making explanations and justifications less called for. Test practitioners utilise the historically forged link between objectivity and quantification and frame contemporary leadership measures as scientific, rational, and superior to qualitative (subjective) evaluation methods. The purportedly superior scientific status of quantitative assessment tools makes questions about the tools' origin, development, and interconnectedness with other tests and psychological theories appear irrelevant.

Further, test practitioners mobilise soft norms about belief, trust, honesty, openness, gratitude, and acceptance, in the attempt to regulate test takers' emotional responses, pre-empt scepticism and critical questions. There is in this way a continuous interplay between the statistical leadership norms within the measures and the soft norms around them. Practitioners intervene in the measurement process in ways that at times promote and protect the statistically derived leadership norms and other times disregard and downplay them. Test practitioners use strategies to highlight certain numbers and discard or moderate others. Through soft norms and emotional regulation efforts, practitioners ensure a certain degree of standardisation of test takers' test experiences and the creation of a compelling and persuasive test personality. When test practitioners mitigate some numbers and norms, while foregrounding others, they create a convincing narrative about the behaviour and/or personality of the test taker, and the areas of the test taker's personality and behaviour that need to be further developed. All these mediating efforts are motivated by test practitioners' knowledge of the unpredictable nature of test takers' varied responses. In other words, test practitioners are aware of test takers' numerous response options, where some counteract the mediating strategies or resist the measures' normalising potentials, prompting test practitioners to employ their strategies.

Based on the above insights, I suggest that the measures do not have their own powerful mathematical agency, meaning that they rarely manage to create the conditions for their own actualisation without external help. Rather, I argue that the measures contain illocutionary and perlocutionary potentials that human force must actualise and turn into effects. If the measures were sufficiently powerful on their own, they would not need to be framed and further 'sold' by test practitioners to the extent that they are.

The measures' performative power thus relies on a particular social context where the test taker has accepted that: 1. I can always do better and improve myself. 2. The test is going to help me with this. 3. The test is trustworthy and reliable. 4. I owe it to the organisation to be open and grateful since it spent a significant amount of money and time on getting me tested. 5. The quicker I accept my results, the better (I am as a leader).

All the work test practitioners put into ensuring these conditions, and test takers' strategies to either accept or resist them, are exactly what deserves more scrutiny.

Implications and future studies

My study of the four measures extends critical literature on quantitative assessment tools by broadening our knowledge of how such tools operate. I contribute with insights into the interaction between social actors and quantitative assessment tools, soft norms and hard norms. I argue that this interaction is key in understanding how quantitative measures succeed or fail in influencing e.g. normalising the attitudes and behaviour of their users.

Based on this understanding, I urge both academics and practitioners to direct their gaze towards the broader context and the social actors involved in a measurement process. Through concepts such as *normalising potentials* and *mediating strategies* we can approach the field of (leadership) measures from a contextual and interactional perspective. Recognising the measures' potentiality means that we foreground the intentions, aims, and agendas underlying them, and what might hinder, obstruct, support, or facilitate them.

This perspective encourages us to look at the strategies social actors employ to either actualise or repress numbers and norms within measures. I argue that we need not only to look at the efforts made to establish such measures as legitimate in the first place, but also to focus on social actors' *continuous* work, since this is essential for the measures' subsequent lives and potential effects. Refocusing our attention to the contextual factors and social actors involved in measurement activities downplays the instruments' abilities, powers, and effects and puts instead the context, influence, and agency of social actors to the centre.

Norms are not universal or fixed. On the contrary, norms are established, meaning that we can negotiate, change, question, and challenge them. This applies to all the norms in and around quantitative measures, which is why I encourage more

questions on the norms present both within and around measurement tools. For example: Why have test developers selected the precise leadership characteristics (norms) they have? What assumptions have informed these choices? Why do scholars and practitioners maintain the conviction that conforming to the norms (in other words, the prescriptive standards) of a measure automatically results in leadership? And what options exist for test takers in responding to norms in and around the measures? Since norms vary and change across context and culture, identifying more norms and strategies used by social actors in relation to different (types of) measures would enrichen our understanding of quantitative tools and the ongoing work they require.

Deeper understandings of the ongoing work tests rely on to realise their normalising power will both make the phenomenon more transparent to users, but also provide scholars and practitioners with arguments that can challenge the instruments' perceived power. Without more transparency, test takers are otherwise left to blindly trust the numbers and their advocates, resulting in both personal and professional costs for the individual. Therefore, and related to past and current calls for an ethics of quantification (e.g. Espeland & Stevens, 2008; Islam, 2022) my study challenges leadership measures' objective status, highlights how this status is connected to ideals of rationality, and draws critical attention to the ways such instruments risk narrowing rather than opening perspectives and (leadership) understandings.

In an emancipatory, ethical quest, knowledge of how social measures work illuminates both the measures' power as well as their weaknesses and shortcomings. Without such insights, the tools' prescriptive standards, labels, and ideas of good and bad leadership can freely flourish, producing simplified understandings about leadership, ourselves, and others. I expect that more studies that take into account the social elements (leadership) measures rely on will further add to knowledge about the instruments' shortcomings, contradictions, and susceptibility to users' counterstrategies, allowing for more critical questioning and perhaps new ways of thinking about (the status and value of) quantitative evaluative tools.

My argument here is not that we must stop measuring leadership or personality altogether; test takers describe positive experiences as well as negative ones (e.g. they report higher self-esteem and more self-awareness). Nevertheless, scholars and practitioners would benefit from a more critical and ethical perspective concerning the tools, an approach which attaches less objectivity and authority to such quantitative measurement activities, affording them less power and a more

limited mandate. In the development of this mindset, I point to the traces of subjectivity, assumptions, paradoxes, and contradictions within the measures, the tools' interconnectivity and concern with normality, and most importantly, the significance of social actors influencing and forming the process, underlining the subjective and normative side of measures and their development. I therefore invite more studies exposing these elements of test use.

If we stop thinking of quantitative measures as indisputably objective, rational, and accurate representations of reality, we might notice the norms and paradoxes in and around them, and reduce the constructed opposition between quantitative measures and subjectivity, rationality and feelings, truth and belief. As shown in the theoretical chapter, objectivity, to which quantification is closely associated, has historically been linked to the negation of subjectivity and thus the obtainment of self-restraint, self-discipline, and self-control (Daston & Galison, 2010). As a result, quantitative tools, presented as objective, carry with them an undermining of the self and of seeing based on 'inference, interpretation, and intelligence' (Daston & Galison, 2010, p.17) – three things that are at the core of human experience. Maintaining the dichotomy between objectivity and subjectivity strengthens the superior status of quantification while problematising qualitative methods and knowledge. This further increases the gap between the notions of objectivity and subjectivity, and the fear of the latter. In the end, this fear of subjectivity reinforces the call for more numbers, more measurements, and objectivity.

If we reduce the gap, by drawing attention to the subjectivity within and around these tools, we can open up for more ways of understanding and approaching leadership (development). In other words, if everything has to pass the test of being put into numbers in order to be considered as important and valid, other interesting, meaningful viewpoints and assessments may be overlooked or disqualified. As Järvinen et al. (2022) argue:

> By framing the conversation around quantified information, an organisation can limit the opportunities of weaker stakeholder groups to participate in a conversation, as those with limited resources are not necessarily able to come up with convincing numbers. (p.24)

Using numbers to evaluate leadership is a way of framing and thereby limiting conversations and viewpoints. Particularly in an area such as leadership, where one could argue that there are as many experiences of leadership as there are people, it might be neither productive nor meaningful to insist on eliminating

subjectivity and achieving the objectivity-imposed ideal of 'blind sight' (Daston & Galison, 2010, p.17). If we accept that leadership and personality are fluid, immeasurable phenomena, we might ease up on the formula hunt, and instead make room for more meaningful ideas about leadership and more varied approaches to evaluating and developing leaders.

Commensurating leadership produces a certain type of knowledge, a knowledge that is shaped by the content and boundaries of the measure. When approaching and describing leadership through quantitative expressions, leadership as a concept and practice appears context-independent, still and fixed (momentarily). When leadership appears stable, this prompts the impression that we have the complete picture of the phenomenon. For example, letting quantification categorise us, dividing us in groups of introverts and extroverts might offer some immediate satisfaction and a sense of deep understanding. However, labelling is not necessarily a conversation starter, but can function as a stopper. Test advocates claim that using tests and their terminology start conversations and aid dialogue, but when we categorise individuals, it can prevent us from asking questions. It is tempting to think we know all there is to know, if we are told someone is an introvert. We don't need to ask further questions or actually get to know the person. On the contrary, one might treat this person according to this classification, in an attempt to respect and accommodate the person's introversion, which further strengthens the perceived legitimacy of the label.

As Porter (1995) argues, quantification can work as a technology of distance. Decision makers can distance themselves from the measured by referring to the truth-telling numbers. Test practitioners and test users often perceive numerically based decisions as less personal, less subjective, and therefore fairer. But such detachment, using numbers as arguments, also comes with a cost. We risk not challenging assumptions, not listening to or ignoring intuitions, emotions, and gut feelings, simply because they are not (yet) expressed numerically. An overreliance on quantitative assessment tools challenges our ability to behave and talk independently from the expectations attached to the tools. In other words, conversations and understandings of leadership are restricted to the terminology and expectations inherent in the measures. In contrast, if we distance ourselves from our numbers, we are forced to stay close to the subject matter, use our words, and think outside the test box.

An overreliance on quantitative assessment tools also carries the possibility of making us more concerned with *seeming* extroverted, empathetic, and ethical, rather than actually *being* it. Continuous tracking and monitoring activities,

resulting in quantitative analyses and categorisations, encourage us to optimise our numbers, reach the target spans, and *appear* better, improved, and extraordinary. Moreover, looking better is only possible because of the in-built comparative mechanism in quantitative leadership and personality tests. Our sense of self-worth or capability is thus tied to comparisons, making social hierarchies more or less directly the foundation for evaluations and decisions.

In conclusion, my thesis informs new approaches to leadership understandings and development by foregrounding the conditions and active interventions which quantitative assessment tools require, their paradoxes and their normative aspects. This suggests that we might benefit from a more cautious, or at least ethically informed reliance on quantitative assessment tools. If we identify norms and ideas of (ab)normality, particular leadership conceptualisations, and paradoxes within leadership measures, and discuss the implicit meanings and assumptions in test items and scale choices, we might initiate interesting and important discussions on what we take for granted, value, and reward, perhaps leading to more nuanced understandings about leadership, how we practice and think about it.

References

Adobe Connect User Community. (n.d.). The Rules of Engagement, Available Online: https://www.connectusers.com/tutorials/2013/06/rules_of_engagement/index.php [Accessed 25 November 2021]

Adolf, M. T. & Stehr, N. (2018). Information, Knowledge, and the Return of Social Physics, *Administration and Society*, vol. 50, no. 9, pp.1238–1258

Alvesson, M. (2011). Interpreting Interviews, SAGE

Alvesson, M. & Kärreman, D. (2016). Intellectual Failure and Ideological Success in Organization Studies: The Case of Transformational Leadership, *Journal of Management Inquiry*, vol. 25, no. 2, pp.139–152

Alvesson, M. & Sköldberg, K. (2009). Reflextive Methodology: New Vistas for Qualitative Research, SAGE

Arthur, W., Woehr, D. J. & Graziano, W. G. (2001). Personality Testing in Employment Settings: Problems and Issues in the Application of Typical Selection Practices, *Personnel Review*, vol. 30, no. 5/6, p.657

Atkinson, P. & Coffey, A. (1997). Analysing Documentary Realities, in *Qualitative Research: Theory, Method and Practice*, SAGE

Austin, J. L. (1962). How to Do Things with Words, London: Oxford University Press

Avolio, B. J. & Gardner, W. L. (2005). Authentic Leadership Development: Getting to the Root of Positive Forms of Leadership, *Leadership Quarterly*, vol. 16, no. 3, pp.315–338

Barbuto, J. E. & Wheeler, D. W. (2006). Scale Development and Construct Clarification of Servant Leadership, *Group and Organization Management*, vol. 31, no. 3, pp.300–326

Barker, R. A. (2001). The Nature of Leadership, *Leadership Perspectives*, vol. 54, no. 4, pp.469–494

Bass, B. M. (1985). Leadership: Good, Better, Best, *Organizational Dynamics*, vol. 13, no. 3, pp.26–40

Bass, B. M. (1990). Bass & Stogdill's Handbook of Leadership - Theory, Research & Managerial Applications, 3rd edn, The Free Press

Bass, B. M. & Avolio, B. J. (1993). Transformational Leadership and Organizational Culture, *Public Administration Quarterly*, vol. 17, no. 1, pp.112–121

Bate, S. P. (1997). Whatever Happened to Organizational Anthropology? A Review of the Field of Organizational Ethnography and Anthropological Studies, *Human Relations*, vol. 50, no. 9, pp.1147–1175

Beer, D. (2016). Metric Power, 1st edn, Palgrave Macmillan

Berger, P. L. & Luckmann, T. (1966). The Social Construction of Reality: A Treatise on the Sociology of Knowledge, Penguin Books

Berman, E. P. & Hirschman, D. (2018). The Sociology of Quantification: Where Are We Now?, *Contemporary Sociology: A Journal of Reviews*, vol. 47, no. 3, pp.257–266

Borsboom, D. (2005). Measuring the Mind, Cambridge University Press

Bowen, C.-C., Martin, B. A. & Hunt, S. T. (2002). A Comparison of Ipsative and Normative Approaches for Ability to Control Faking in Personality Questionnaires, *The International Journal of Organizational Analysis*, vol. 10, no. 3, pp.240–259

Brett, J. F. & Atwater, Leanne, E. (2001). 360° Feedback: Accuracy, Reactions, and Perceptions of Usefulness, *Jornal of Applied Psychology*, vol. 86, no. 5, pp.930–942

Bruno, I., Jany-Catrice, F. & Touchelay, B. (2016). The Social Sciences of Quantification: From Politics of Large Numbers to Target-Driven Policies, 1st edn, Springer

Bryman, A. (1986). Leadership and Organizations, Routledge & Kegan Paul

Butcher, J. N. (2010). Personality Assessment from the Nineteenth to the Early Twenty-First Century: Past Achievements and Contemporary Challenges, *Annual Review of Clinical Psychology*, vol. 6, no. 1, pp.1–20

Carlyle, T. (1840). On Heroes, Hero-Worship and the Heroic in History, London: Chapman and Hall

Cattell, B. Y. R. B. (1944). Psychological Measurement: Normative, Ipsative, Interactive, *Psychological Review*, pp.292–303

Chan, D., Schmitt, N., Sacco, J. M. & DeShon, R. P. (1998). Understanding Pretest and Posttest Reactions to Cognitive Ability and Personality Tests, *Journal of Applied Psychology*, vol. 83, no. 3, pp.471–485

Charmaz, K. (2006). Constructing Grounded Theory - A Practical Guide Through Qualitative Analysis, SAGE

Cicourel, A. Vi. (1964). Method and Measurement in Sociology, Free Press of Glencoe

Ciulla, J. B. (2004). Ethics, the Heart of Leadership, Praeger

Cohen, P. C. (1999). A Calculating People - The Spread of Numeracy in Early America, New York and London: Routledge

Collins Dictionary. (n.d.). Definition of Norm, Available Online: https://www.collinsdictionary.com/dictionary/english/norm [Accessed 22 March 2018]

Committee on the U.S. Naturalization Test Redesign. (2004). Redesigning the U.S. Naturalization Tests: Interim Report, Washington, DC: The National Academies Press

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R. & Butera, H. (2008). Comparing Personality Test Formats and Warnings: Effects on Criterion-Related Validity and Test-Taker Reactions, *International Journal of Selection and Assessment*, vol. 16, no. 2, pp.155–169

Cooper, R. D. & Schindler, S. P. (2014). Business Research Methods, 12th edn, McGraw Hill

Cooren, F. (2004). Textual Agency: How Texts Do Things in Organizational Settings, *Organization*, vol. 11, no. 3, pp.373–393

Crosby, A. W. (1997). The Measure of Reality - Quantification and Western Society, 1250-1600, Cambridge University Press

Dalakoura, A. (2010). Differentiating Leader and Leadership Development, *Journal of Management Development*, vol. 29, no. 5, pp.432–441

Danziger, K. (1990). Constructing the Subject: Historical Origins of Psychological Research, Cambridge University Press

Daston, L. & Galison, P. (2010). Objectivity, Zone Books

Datta, B. (2015). Assessing the Effectiveness of Authentic Leadership, *International Journal of Leadership Studies*, vol. 9, no. 1, pp.62–75

Davis, M. S. (1971). That's Interesting! Towards a Phenomonology of Sociology and a Sociology of Phenomonology, *Philosophy of Social Sciences*, vol. 1, no. 2, pp.309–344

Denzin, N. K. & Lincoln, Y. S. (2000). The Discipline and Practice of Qualitative Research, in *Handbook of Qualitative Research*, Thousand Oaks: SAGE

Desrosières, A. (1998). The Politics of Large Numbers - A History of Statistical Reasoning, *Harvard University Press*

Dorn, C. (2019). When Reactivity Fails: The Limited Effects of Hospital Rankings, *Social Science Information*, vol. 58, no. 2, pp.327–353

Einola, K. & Alvesson, M. (2020). Behind the Numbers: Questioning Questionnaires, *Journal of Management Inquiry*, vol. 30, no. 1, pp.1–13

Elmes, M. B. & Costello, M. (1992). Mystification and Social Drama: The Hidden Side of Communication Skills Training, *Human Relations*, vol. 45, no. 5, pp.427–445

Espeland, W. N. & Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds, *American Journal of Sociology*, vol. 113, no. 1, pp.1–40

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

Espeland, W. N. & Stevens, M. L. (2008). A Sociology of Quantification, *European Journal of Sociology*, vol. 49, no. 3, pp.401–436

Fauré, B., Cooren, F. & Matte, F. (2019). To Speak or Not to Speak the Language of Numbers: Accounting as Ventriloquism, *Accounting, Auditing and Accountability Journal*, vol. 32, no. 1, pp.337–361

Ferry, N. & Guthey, E. (2021). Start 'Em Early: Pastoral Power and the Confessional Culture of Leadership Development in the US University, *Journal of Business Ethics*, vol. 173, no. 4, pp.723–736

Flyvbjerg, B. (2006). Five Misunderstandings about Case-Study Research, *Qualitative Inquiry*, pp.219–245

Fontana, A. & Frey, J. H. (2000). The Interview: From Structured Quetions to Negotiated Text, in *Handbook of Qualitative Research*, Thousand Oaks: SAGE

Foucault, M. (1991). Discipline and Punish - The Birth of the Prison, Penguin Books

Fry, L. W. (2003). Toward a Theory of Spiritual Leadership, *Leadership Quarterly*, vol. 14, no. 6, pp.693–727

Gagnon, S. & Collinson, D. (2014). Rethinking Global Leadership Development Programmes: The Interrelated Significance of Power, Context and Identity, *Organization Studies*, vol. 35, no. 5, pp.645–670

Gibby, R. E. & Zickar, M. J. (2008). A History of the Early Days of Personality Testing in American Industry: An Obsession with Adjustment, *History of Psychology*, vol. 11, no. 3, pp.164–184

Gioia, D. A., Corley, K. G. & Hamilton, A. L. (2013). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology, *Organizational Research Methods*, vol. 16, no. 1, pp.15–31

Global HR Research. (n.d.). Candidate Screening: The Benefits of Personality Testing, Available Online: https://www.ghrr.com/candidate-screening-the-benefits-of-personality-testing/ [Accessed 7 January 2020]

Gould, S. J. (1996). The Mismeasure of Man, W. W. Norton & Company

Greenthumbs. (2022). Dismantling Common Myths About Personality Tests, Available Online: https://blog.greenthumbs.in/personality-tests-myths.php [Accessed 7 January 2020]

Gregory, S. (2013). The Tabulation of England: How the Social World Was Brought in Rows and Columns, *Distinktion*, vol. 14, no. 3, pp.305–325

Hacking, I. (1990). The Taming of Chance, *History of European Ideas*, Cambridge University Press

Hacking, I. (2006). Making Up People, *London Review of Books*, Available Online: https://www.lrb.co.uk/the-paper/v28/n16/ian-hacking/making-up-people [Accessed 9 March 2022]

Hanson, N. R. (2000). Observation, in *Readings in the Philosophy of Science: From Positivism to Postmodernism*, Mayfield Pub.

Hicks, L. E. (1970). Some Properties of Ipsative, Normative, and Forced-Choice Normative Measures, *Psychological Bulletin*, vol. 74, no. 3, pp.167–184

Hogan Assessments. (n.d.). About Hogan, Available Online: https://www.hoganassessments.com/about/ [Accessed 9 December 2019]

Hogan Assessments. (n.d.). Talent Development, Available Online: https://www.hoganassessments.com/talent-development [Accessed 3 March 2020]

Hogan Assessments. (n.d.). Leadership Forecast Series, Available Online: https://www.hoganassessments.com/products/leadership-forecast-series/ [Accessed 20 December 2021]

Hogan Assessments. (n.d.). Turn Leadership into an Exact Science, Available Online: https://www.hoganassessments.com/talent-development/leadership-development/ [Accessed 20 December 2021]

Human Developers. (n.d.). About Us, Available Online: https://humandevelopers.com/en/om-os/ [Accessed 6 January 2023]

Human Developers. (n.d.) Håndbog for Human Developers Test System

Islam, G. (2022). Business Ethics and Quantification: Towards an Ethics of Numbers, *Journal of Business Ethics*, vol. 176, no. 2, pp.195–211

Islam, G. & Greenwood, M. (2022). The Metrics of Ethics and the Ethics of Metrics, *Journal of Business Ethics*, vol. 175, no. 1, pp.1–5

Ivanov, S., McFadden, M. & Anyu, J. N. (2021). Examining and Comparing Good and Bad Leaders Based on Key Leadership Characteristics: A Leadership Case Study., *International Journal of Organizational Innovation*, vol. 13, no. 3, pp.275–281

Jones, G. (2007). Of Transactional And Transformational Leadership – Question Of Balance, *Journal of Business*, vol. 5, no. 11, pp.1–8

Kane, G. C., Gotto, J. L., Mangione, S., West, S. & Hojat, M. (2007). Jefferson Scale of Patient's Perceptions of Physician Empathy: Preliminary Psychometric Data, *Croatian Medical Journal*, vol. 48, no. 1, pp.81–86

Kline, P. (1988). Psychology Exposed - Or The Emperor's New Clothes, Routledge

Kline, P. (1998). The New Psychometrics - Science, Psychology and Measurement, Routledge

Kvale, S. (1994). Ten Standard Objections to Qualitative Research Interviews, *Journal of Phenomenological Psychology*, vol. 25, no. 2, pp.147–173

Lashway, L. (1997). Measuring Leadership Potential, *ERIC Digest*, [e-journal], Available Online: https://www.ericdigests.org/1998-1/measuring.htm [Accessed 7 February 2018]

Lashway, L. (1998). Measuring Leadership, *Research roundup*, vol. 14, no. 2, pp.2–5

Lazarsfeld, P. F. (1961). Notes on the History of Quantification in Sociology - Trends, Sources and Problems, *The University of Chicago Press on behalf of The History of Science Society*, vol. 52, no. 2, pp.277–333

Management Research Group (2010). Leadership Effectiveness Analysis™: Technical Considerations Report

Leskiw, S.-L. & Singh, P. (2007). Leadership Development: Learning from Best Practices, *Leadership & Organization Development Journal*, vol. 28, no. 5, pp.444–464

Lincoln, Y. S. & Guba, E. G. (1985). Naturalistic Inquiry, SAGE

Locke, W. (2014). The Intensification of Rankings Logic in an Increasingly Marketised Higher Education Environment, *European Journal of Education*, vol. 49, no. 1, pp.77–90

Lord, R. G. & Hall, R. J. (2005). Identity, Deep Structure and the Development of Leadership Skill, *Leadership Quarterly*, vol. 16, no. 4, pp.591–615

Maanen, J. Van. (1979). The Fact of Fiction in Organizational Ethnography, *Administrative Science Quarterly*, vol. 24, no. 4, pp.539–551

Mabey, C. (2013). Leadership Development in Organizations: Multiple Discourses and Diverse Practice, *International Journal of Management Reviews*, vol. 15, no. 4, pp.359–380

Mau, S. (2019). The Metric Society - On the Quantification of the Social, Polity Press

Mehrabani, S. E. & Mohamad, N. A. (2015). New Approach to Leadership Skills Development (Developing a Model and Measure), *Journal of Management Development*, vol. 34, no. 7, pp.821–853

Meier, F. & Carroll, B. (2019). Making up Leaders: Reconfiguring the Executive Student through Profiling, Texts and Conversations in a Leadership Development Programme, *Human Relations,* vol. 1, no. 12, pp.1-23

Melamed, T. & Jackson, D. (1995). Psychometric Instruments: Potential Benefits and Practical Use, *Industrial and Commercial Training*, vol. 27, no. 4, pp.11–16

Mennicken, A. & Espeland, W. N. (2019). What's New with Numbers? Sociological Approaches to the Study of Quantification, *Annual Review of Sociology*, vol. 45, pp.223–245

Mennicken, A. & Salais, R. (eds). (2022). The New Politics of Numbers - Utopia, Evidence and Democracy, Palgrave Macmillan

Merriam-Webster. (n.d.). Norm, Available Online: https://www.merriam-webster.com/dictionary/norm [Accessed 22 March 2018]

Messick, S. (1980). Test Validity and the Ethics of Assessment, *American Psychologist*, vol. 35, no. 11, pp.1012–1027

Mills, J. P. & Boardley, I. D. (2017). Development and Initial Validation of an Indirect Measure of Transformational Leadership Integrity, *Psychology of Sport and Exercise*, vol. 32, pp.34–46

MindTools. (2022). How Good Are Your Leadership Skills?, Available Online: https://www.mindtools.com/pages/article/newLDR_50.htm [Accessed 6 February 2018]

Moore, P. & Robinson, A. (2015). The Quantified Self: What Counts in the Neoliberal Workplace, *New Media & Society*, vol. 18, no. 11, pp.2774–2792

Morgan, G. (1980). Paradigms, Metaphos, and Puzzle Solving in Organization Theory, *Administrative Science Quaterly*, vol. 25, no. 4, pp.605–622

Muller, J. Z. (2019). The Tyranny of Metrics, Princeton University Press

Ni, Y. & Hauenstein, N. M. A. (1998). Applicant Reactions to Personality Tests: Effects of Item Invasiveness and Face Validity, *Journal of Business and Psychology*, vol. 12, no. 4, pp.391–406

Nicholson, H. & Carroll, B. (2013). Identity Undoing and Power Relations in Leadership Development, *Human Relations*, vol. 66, no. 9, pp.1225–1248

Nussbaum, M. C. & Hursthouse, R. (1984). Plato on Commensurability and Desire, *Oxford University Press on behalf of The Aristotelian Society*, vol. 58, pp.55–96

People Test Systems. (n.d.). Overview of Test Tools, Available Online: https://www.peopletestsystems.com/test-tools [Accessed 10 December 2019]

Porter, T. (2012). Funny Numbers, *Culture Unbound,* vol. 4, pp.585–598

Porter, T. M. (1995). Trust in Numbers - The Pursuit of Objectivity in Science and Public Life, Princeton University Press

Prasad, P. (2005). Crafting Qualitative Research: Working in the Postpositivist Traditions, M.E. Sharpe

Psychometric Success. (2022). Why Do Employers Use Psychometric Tests, Available Online: http://www.psychometric-success.com/psychometric-tests/psychometric-tests-introduction.htm [Accessed 22 March 2018]

Psychometric Tests. (2013). Leadership Test, Available Online: https://www.psychometrictest.org.uk/leadership-test/ [Accessed 11 July 2022]

Rani, J. S. (2004). Educational Measurement and Evaluation, Discovery Publishing House

Rettberg, J. W. (2014). Seeing Ourselves Through Technology, 2nd edn, Palgrave Macmillan

Robinson, M. A. (2017). Using Multi-Item Psychometric Scales for Research and Practice in Human Resource Management, *Human Resource Management*, vol. 57, no. 3, pp.739–750

Rose, N. (2008). Psychology as a Social Science, *Subjectivity*, vol. 25, no. 1, pp.446–462

Ryan, G. W. & Bernard, H. R. (2003). Techniques to Identify Themes, *Field Methods*, vol. 15, no. 1, pp.85–109

Schriesheim, C. A. & Bird, B. J. (1979). Contributions of the Ohio State Studies to the Field of Leadership, *Journal of Management*, vol. 5, no. 2, pp.135–145

Schriesheim, C. & Kerr, S. (1974). Psychometric Properties of the Ohio State Leadership Scales, *Psychological Bulletin*, vol. 81, no. 11, pp.756–765

Shamir, B. & Eilam, G. (2005). 'What's Your Story?' A Life-Stories Approach to Authentic Leadership Development, *Leadership Quarterly*, vol. 16, no. 3, pp.395–417

Sigma Assessment Systems. (n.d.). Leadership Character Insight Assessment, Available Online: http://www.sigmaassessmentsystems.com/assessments/leadership-character-insight-assessment/ [Accessed 6 February 2018]

Spector, B. A. (2016). Carlyle, Freud, and the Great Man Theory More Fully Considered, *Leadership*, vol. 12, no. 2, pp.250–260

Spoelstra, S. (2018). Leadership and Organization - A Philosophical Introduction, Routledge

Stake, R. (2000). Case Studies, in *Handbook of Qualitative Research*, Thousand Oaks: SAGE

Stake, R. (2010). Qualitative Research - How Things Work, in *Qualitative Research - Studying How Things Work*, Guilford Publications

Stein, Z. (2016). Social Justice and Educational Measurement, Routledge

Swedberg, R. (2012). Theorizing in Sociology and Social Science: Turning to the Context of Discovery, *Theory and Society*, vol. 41, no. 1, pp.1–40

T. Järvinen, J., Laine, M., Hyvönen, T. & Kantola, H. (2022). Just Look at the Numbers: A Case Study on Quantification in Corporate Environmental Disclosures, *Journal of Business Ethics*, vol. 175, no. 1, pp.23–44

The Extraordinary Leader - Participant Manual. (2015)

Thomas. (2022). 360 Degree Feedback Assessment, Available Online: https://www.thomas.co/assessments/360-degree-feedback [Accessed 7 January 2020]

Townley, B. (1993). Foucault, Power / Knowledge, and It's Relevance for Human Resource Management, *Academy of Management Review*, vol. 18, no. 3, pp.518–546

Unique HR. (2016). Are You Conducting Leadership Skills Assessments? Here's Why You Should Be, Available Online: https://www.uniquehr.com/are-you-conducting-leadership-skills-assessments-heres-why-you-should-be/ [Accessed 6 February 2018]

Van Dusen, A. C. (1948). Measuring Leadership Ability, *Personnel Psychology*, vol. 1, no. 1, pp.67–79

Vincent, J. (2022). Made to Measure: Why We Can't Stop Quantifying Our Lives, *The Guardian*, Available Online: https://www.theguardian.com/news/2022/may/26/measurement-why-we-cant-stop-quantifying-our-lives [Accessed 16 June 2022]

Visser, R. & Schaap, P. (2017). Job Applicants' Attitudes towards Cognitive Ability and Personality Testing, *SA Journal of Human Resource Management*, vol. 15, pp.1–11

Walumbwa, F. O., Avolio, B. J., Gardner, W. L., Wernsing, T. S. & Peterson, S. J. (2008). Authentic Leadership: Development and Validation of a Theory-Based Measure, *Journal of Management*, vol. 34, no. 1, pp.89–126

Weber, M. (1968). On Charisma and Institution Building: Selected Papers, edited by S. N. Eisenstadt, Chicago, IL: University of Chicago Press

Weber, M. (1978). Charismatic Authority, in *Economy and Society: An Outline of Interpretative Sociology*

Wilson, S., Lee, H., Ford, J. & Harding, N. (2020). On the Ethics of Psychometric Instruments Used in Leadership Development Programmes, *Journal of Business Ethics*, vol. 172, no. 2, pp.211–227

Wood, J. D. & Petriglieri, G. (2004). The Merchandising of Leadership, in S. Chowdhury (ed.), *Next Generation: New Strategies from Tomorrow's Thought Leaders*, NJ, USA: John Wiley & Sons, pp.200–219

Zenger Folkman. (2021). The Extraordinary Leader, Available Online: https://zengerfolkman.com/360-degree-assessment/ [Accessed 17 December 2021]

Zenger, J. & Folkman, J. The Extraordinary Leader, Available Online: https://zengerfolkman.com/wp-content/uploads/2022/04/Extraordinary-Leader-2nd-Edition-Product-Overview_020722.pdf

Zenger, J. & Folkman, J. (2017). Key Insights From the Extraordinary Leader, Available Online: https://zengerfolkman.com/wp-content/uploads/2019/05/White-Paper-Extraordinary-Leader-Insights-Excerpts-from-The-Extraordinary-Leader.pdf [Accessed 7 March 2023]

Zenger, J., Folkman, J. & Sandholtz, K. (2021). Leadership under the Microscope, Available Online: https://zengerfolkman.com/wp-content/uploads/2022/04/Leadership-Under-the-Microscope_01.21.21.pdf

Zenger, J. H. & Folkman, J. R. (2012). How to Be Exceptional: Drive Leadership Success by Magnifying Your Strengths, McGraw Hill

APPENDICES

Appendix 1: Overview of competencies, items, and behaviour (The Extraordinary Leader)

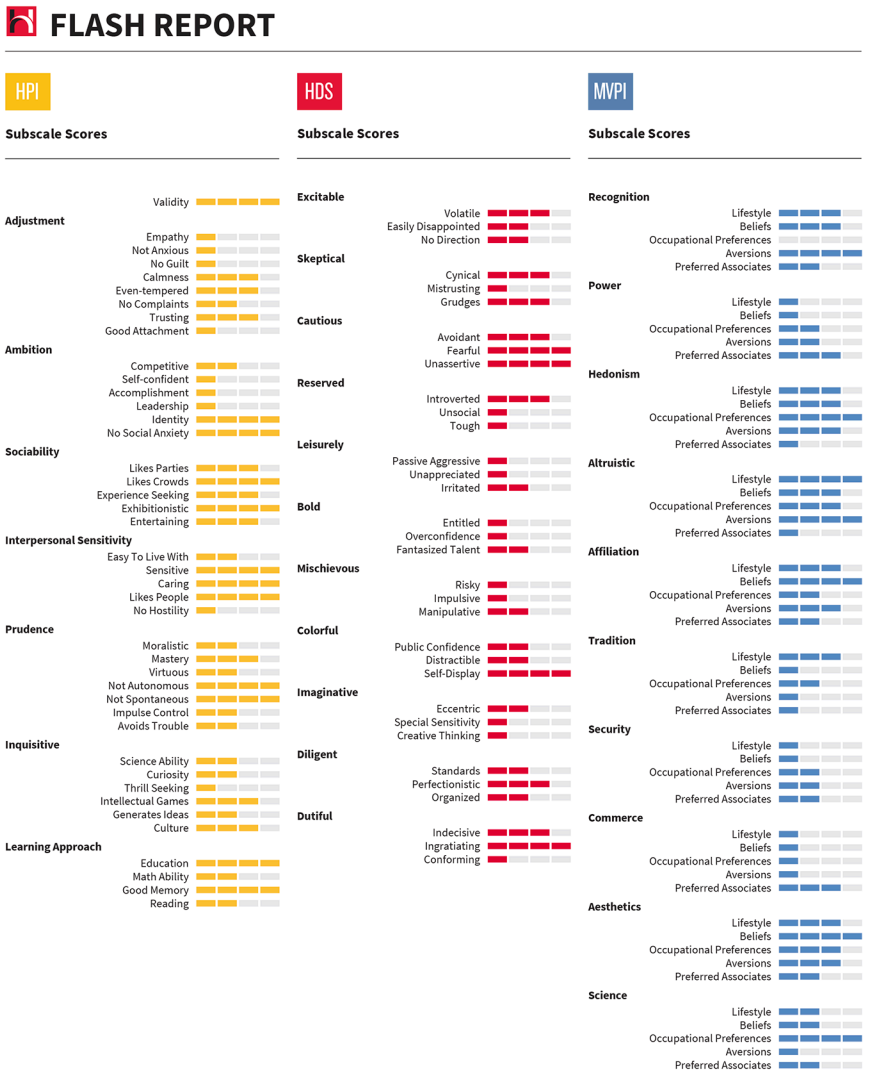
Competency	Items	Behaviour
1. Displays high integrity and honesty	1. Is a role model and sets a good example for his/her work group 2. Works hard to "walk the talk" and avoids saying one thing and doing another 3. Is careful to honor commitments and keep promises	**Character**
2. Technical/Professional Expertise	4. Many people seek after his/her opinions 5. His/her skills and knowledge make an important contribution to achieving team results 6. Teammates trust his/her ideas and opinions because of in-depth knowledge and experience	**Personal Capability**
3. Solves Problems and Analyzes Issues	7. Has the ability to anticipate and respond quickly to problems 8. Is trusted by others to use good judgment when making decisions 9. Spots new trends, potential problems, and opportunities early	**Personal Capability**

4. Innovates	10. Frequently encourages others to consider new approaches and ideas (e.g. avoids getting struck in a "one right way" approach") 11. Finds ways to improve new ideas rather than discourage them 12. Constructively challenges the standard approaches and finds important processes to get work done	**Personal Capability**
5. Practices Self-Development	13. Makes a real effort to improve based on feedback from others 14. Actively looks for opportunities to get feedback to improve him/herself 15. Creates an atmosphere of continual improvement in which self and others push to exceed the expected results	**Personal Capability**
6. Drives for Results	16. Does everything possible to achieve goals 17. Achieves agreed-upon goals within the time allotted 18. Follows through on objectives to ensure successful completion i.e. does NOT get distracted before project is completed	**Focus on Results**
7. Establishes Stretch Goals	19. Establishes high standards of excellence for the work group 20. Is skilful at getting people to stretch for goals that go beyond what they originally thought possible 21. Keeps people focused on the highest priority goals and objectives	**Focus on Results**
8. Takes initiative	22. Can always be counted on to follow through on commitments 23. Willingly goes above and beyond what needs to be done 24. Is energized and excited to take on challenging goals, for which	**Focus on Results**

	he/she is held personally accountable	
9. Communicates Powerfully and Prolifically	25. Provides others with a definite sense of direction and purpose 26. Skilled at communicating insights and understanding of issues or problems 27. Helps people understand how their work contributes to broader business objectives	**Interpersonal Skills**
10. Inspires and Motivates Others to High Performance	28. Energizes people to achieve exceptional results 29. Inspires others to high level of effort and performance 30. Brings to the group a high level of energy and enthusiasm	**Interpersonal Skills**
11. Builds Relationships	31. Balances "getting results" with a concern for others' needs 32. Is trusted by all members of the work group 33. Stays on touch with issues and concerns of individuals in the work group	**Interpersonal Skills**
12. Develops Others	34. Provides coaching and acts as a mentor to others 35. Is truly concerned about developing others 36. Gives honest feedback in a helpful way	**Interpersonal Skills**
13. Collaboration and Teamwork	37. Promotes high level of cooperation between all members if the work group 38. Resolves conflict within the work group 39. Achieves objectives requiring a high level of cooperation from people in other parts of the organization	**Interpersonal Skills**
14. Develops Strategic Perspective	40. Helps others understand the organization's vision and objectives so that they can translate them into challenging and meaningful goals	**Leading Change**

	41. Maintains a clear perspective between the overall picture and the details 42. Has a perspective beyond the "day-to-day" work to take a longer-term, broader view of business decisions	
15. Champions Change	43. Quickly recognizes situations where change is needed 44. Is willing to become a champion for new projects or programs, presenting them so that others support them 45. Does an excellent job of marketing projects, programs or products 46. Has the courage to make the change that will improve the organization	**Leading Change**
16. Connects the Group to the Outside World	47. Helps people understand how meeting customers' needs is central to the mission and goals of the organization 48. Has demonstrated ability to represent the organization to key groups 49. Is the antenna for the organization, bringing in relevant information that benefits the group	**Leading Change**

Appendix 2: Overview of scales and subscales (Hogan Leadership Forecast, own report)



Lund Studies in Economics and Management

Editor, issues 157–	Niclas Andrén
Editor, issues 142–156	Charlotta Levay
Editor, issues 109–141	Thomas Kalling
Editors, issues 88–108	Mats Benner & Thomas Kalling
Editor, issues 1–87	Allan T. Malm

166. Emilie Hesselbo (2023): *Sizing up leadership – norms and normativity in and around leadership measures*

165. Erik Brattström (2023): *The missing link: The implementation of priorities for research, development, and innovation*

164. Axel Welinder (2023): *Legitimizing sustainability talk in retail talk – The case of IKEAs sustainability journey*

163. Pelle Högnelid (2022): *Purposeful Combination: Management of Knowledge Integration in the Development of Self-Driving Cars*

162. Hossain Shahriar (2022): *Gender Transculturation: Navigating Market-Mediated Contesting Gender Ideologies in Consumer Acculturation*

161. Johan Gromark (2022): *Brand orientation in action – Towards a relational approach*

160. Henrik Edlund (2022): *Organizational and individual response to hybridity in the public sector: A case study exploring the customer orientation of the Swedish Enforcement Authority*

159. Karin Alm (2022): *Butikens roll för hållbar konsumtion – Ett marknadsorienterat perspektiv på hållbar sortimentsutveckling i dagligvaruhandeln*

158. Jayne Jönsson (2022): *Logic Salience: Navigating in the institutional landscape of funding volatility and ideological disputes in nonprofit hybrid organizing*

157. Jonas Cedergren (2022): *Becoming a Physician-Scientist: A Study on the Power of Membership in Communities of Practice*

156. Wenjun Wen (2021): *Rethinking Accounting Professionalisation in China: A Study of the Development of the Chinese Public Accounting Profession since the "Reform and Opening-up"*

155. Parfait Yongabo (2021): *Fostering Knowledge uptake in Emerging Innovation Systems: Enhancing Conditions for Innovation in Rwanda*

154. Maria Bengtsson (2021): *National adoption of International Financial Reporting Standards: The case of China*

153. Janina Schaumann (2021): *Stakeholder-based brand equity (SBBE) – A qualitative study of its development through firm-stakeholder interactions in emerging markets*

152. Anna Stevenson (2021): *Constructing the 'social' in social entrepreneurship: A postcolonial perspective*

151. Tanya Kolyaka (2021): *Financial Bootstrapping as Relational Contract: Linking resource needs, bootstrapping behaviors, and outcomes of bootstrapping exchanges*

150. Louise Klintner (2021): *Normalizing the Natural: A study of menstrual product destigmatization*

149. Zahida Sarwary (2019): *Puzzling out the choice of capital budgeting techniques among high-growth small and medium sized firms*

148: Vivek Kumar Sundriyal (2019): *Entrepreneurship as a career: An investigation into the pre-entrepreneurship antecedents and post-entrepreneurship outcomes among the Science and Technology Labor Force (STLF) in Sweden*

147. Ziad El-Awad (2019): *Beyond individuals – A Process of Routinizing Behaviors Through Entrepreneurial Learning: Insights from Technology-Based Ventures*

146. Carys Egan-Wyer (2019): *The Sellable Self: Exploring endurance running as an extraordinary consumption experience*

145. Lisa Källström (2019): *'A good place to live' – Residents' place satisfaction revisited*

144. Anamaria Cociorva (2019): *Essays on Credit Ratings*

143. Elisabeth Kjellström (2019): *Outsourcing of Organizational Routines: Knowledge, control, and learning aspects*

142. Erik Ronnle (2019): *Justifying Mega-Projects: An Analysis of the Swedish High-Speed Rail Project*

141. Gustav Hägg (2017): *Experiential entrepreneurship education: Reflective thinking as a counterbalance to action for developing entrepreneurial knowledge*

140. Mathias Skrutkowski (2017): *Disgraced. A study of narrative identity in organizations that suffer crises of confidence*

139. Ana Paula do Nascimento (2017): *Funding matters: A study of internationalization programs in science, technology and innovation*

138. Amalia Foukaki (2017): *Corporate Standardization Management: A Case Study of the Automotive Industry*

137. Nathalie Larsson (2016): *From performance management to managing performance: An embedded case study of the drivers of individual and group-based performance in a call center context*

136. Clarissa Sia-Ljungström (2016): *Connecting the Nodes – An interactive perspective on innovative microenterprises in a mature industry*

135. Sten Bertil Olsson (2016): *Marknadsreglering och dess effekter på regionala och lokala gymnasiemarknaders funktion*

134. Mattias Haraldsson (2016): *Accounting choice, compliance and auditing in municipal organisations*

133. Kaj-Dac Tam (2016): *Perceptual Alignment of Retail Brand Image in Corporate Branding: A study of employee perceived stakeholder alignment and effects on brand equity*

132. Wen Pan-Fagerlin (2016): *Participant, Catalyst or Spectator? A study of how managers apply control in innovation processes*

131. Yaqian Wang (2014): *Inside the Box – Cultures of Innovation in a Mature Industry*

130. Paul Pierce (2013): *Using Alliances to Increase ICT Capabilities*

129. Linn Andersson (2013): *Pricing capability development and its antecedents*

128. Lena Hohenschwert (2013): *Marketing B2B Sales Interactions Valuable – A Social and Symbolic Perspective*

127. Pia Nylinder (2012): *Budgetary Control in Public Health Care – A Study about Perceptions of Budgetary Control among Clinical Directors*

126. Liliya Altshuler (2012): *Competitive Capabilities of a Technology Born Global*

125. Timurs Umans (2012): *The bottom line of cultural diversity at the top – The top management team's cultural diversity and its influence on organizational outcomes*

124. Håkan Jankensgård (2011): *Essays on Corporate Risk Management*

123. Susanne Lundholm (2011): *Meta-managing – A Study on How Superiors and Subordinates Manage Their Relationship in Everyday Work*

122. Katarzyna Cieślak (2011): *The Work of the Accounting & Controlling Department and its Drivers: Understanding the concept of a business partner*

121. Ulf Elg, Karin Jonnergård (editors): *Att träda in i en profession: Om hur kvinnor och män etablerar sig inom revisionsbranschen och akademin*

120. Jonas Fjertorp (2010): *Investeringar i kommunal infrastruktur – Förutsättningar för en målfokuserad investeringsverksamhet*

119. Fredrik Ericsson (2010): *Säkringsredovisning – Implementeringen av IAS 39 i svenska icke-finansiella börsföretag och konsekvenser för säkringsverksamheten*

118. Steve Burt, Ulf Johansson, Åsa Thelander (editors, 2010): *Consuming IKEA. Different perspectives on consumer images of a global retailer*

117. Niklas Persson (2010): *Tracing the drivers of B2B brand strength and value*

116. Sandra Erntoft (2010): *The use of health economic evaluations in pharmaceutical priority setting – The case of Sweden*

115. Cecilia Cassinger (2010): *Retailing Retold – Unfolding the Process of Image Construction in Everyday Practice*

114. Jon Bertilsson (2009): *The way brands work – Consumers' understanding of the creation and usage of brands*

113. Per Magnus Andersson, Peter Jönsson, Gert Paulsson, Stefan Yard (editors, 2009): *Ett smörgåsbord med ekonomistyrning och redovisning – En vänbok till Olof Arwidi*

112. Agneta Moulettes (2009): *The discursive construction, reproduction and continuance of national cultures – A critical study of the cross-cultural management discourse*

111. Carl Cederström (2009): *The Other Side of Technology: Lacan and the Desire for the Purity of Non-Being*

110. Anna Thomasson (2009): *Navigating in the landscape of ambiguity – A stakeholder approach to the governance and management of hybrid organisations*

109. Pia Ulvenblad (2009): *Growth Intentions and Communicative Practices – Strategic Entrepreneurship in Business Development*

108. Jaqueline Bergendahl (2009): *Entreprenörskapsresan genom beslutsprocesser i team – En elektronisk dagboksstudie i realtid*

107. Louise D. Bringselius (2008): *Personnel resistance in mergers of public professional service mergers – The merging of two national audit organizations*

106. Magnus Johansson (2008): *Between logics – Highly customized deliveries and competence in industrial organizations*

105. Sofia Avdeitchikova (2008): *Close-ups from afar: the nature of the informal venture capital market in a spatial context*

104. Magnus Nilsson (2008): *A Tale of Two Clusters – Sharing Resources to Compete*

103. Annette Cerne (2008): *Working with and Working on Corporate Social Responsibility: The Flexibility of a Management Concept*

102. Sofia Ulver-Sneistrup (2008): *Status Spotting – A Consumer Cultural Exploration into Ordinary Status Consumption of "Home" and Home Aesthetics*

101. Stefan Henningsson (2008): *Managing Information Systems Integration in Corporate Mergers and Acquisitions*

100. Niklas L. Hallberg (2008): *Pricing Capability and Its Strategic Dimensions*

99. Lisen Selander (2008): *Call Me Call Me for Some Overtime – On Organizational Consequences of System Changes*

98. Viktorija Kalonaityte (2008): *Off the Edge of the Map: A Study of Organizational Diversity as Identity Work*

97. Anna Jonsson (2007): *Knowledge Sharing Across Borders – A Study in the IKEA World*

96. Sverre Spoelstra (2007): *What is organization?*

95. Veronika Tarnovskaya (2007): *The Mechanism of Market Driving with a Corporate Brand – The Case of a Global Retailer*

94. Martin Blom (2007): *Aktiemarknadsorienteringens ideologi – En studie av en organisations försök att skapa aktieägarvärde, dess styrning och kontroll samt uppgörelse med sitt förflutna*

93. Jens Rennstam (2007): *Engineering Work – On Peer Reviewing as a Method of Horizontal Control*

92. Catharina Norén (2007): *Framgång i säljande – Om värdeskapande i säljar- och köparinteraktionen på industriella marknader*

91. John Gibe (2007): *The Microstructure of Collaborative E-business Capability*

90. Gunilla Nordström (2006): *Competing on Manufacturing – How combinations of resources can be a source of competitive advantage*

89. Peter W Jönsson (2006): *Value-based management – positioning of claimed merits and analysis of application*

88. Niklas Sandell (2006): *Redovisningsmått, påkopplade system och ekonomiska konsekvenser – Redovisningsbaserade prestationsersättningar*

87. Nadja Sörgärde (2006): *Förändringsförsök och identitetsdramatisering. En studie bland nördar och slipsbärare*

86. Johan Alvehus (2006): *Paragrafer och profit. Om kunskapsarbetets oklarhet*

85. Paul Jönsson (2006): *Supplier Value in B2B E-Business – A case Study in the Corrugated Packaging Industry*

84. Maria Gårdängen (2005): *Share Liquidity and Corporate Efforts to Enhance it – A study on the Swedish Stock Exchange*

83. Johan Anselmsson, Ulf Johansson (2005): *Dagligvaruhandelns egna varumärken – konsekvenser och utvecklingstendenser*

82. Jan Alpenberg, Fredrik Karlsson (2005): *Investeringar i mindre och medelstora tillverkande företag - drivkrafter, struktur, process och beslut*
81. Robert Wenglén (2005): *Från dum till klok? – en studie av mellanchefers lärande*
80. Agneta Erfors (2004): *Det är dans i parken ikväll – Om samverkan mellan näringsliv och akademi med forskningsparken som mäklande miljö och aktör*
79. Peter Svensson (2004): *Setting the Marketing Scene. Reality Production in Everyday Marketing Work*
78. Susanne Arvidsson (2003): *Demand and Supply of Information on Intangibles: The Case of Knowledge-Intense Companies*
77. Lars Nordgren (2003): *Från patient till kund. Intåget av marknadstänkande i sjukvården och förskjutningen av patientens position*
76. Marie Löwegren (2003): *New Technology Based Firms in Science Parks. A Study of Resources and Absorbtive Capacity*
75. Jacob Östberg (2003): *What's Eating the Eater? Perspectives on the Everyday Anxiety of Food Consumption in Late Modernity*
74. Anna Stafsudd (2003): *Measuring the Unobservable: Selecting Which Managers for Higher Hierarchical Levels*
73. Henrick Gyllberg, Lars Svensson (2002): *Överensstämmelse mellan situationer och ekonomistyrsystem – en studie av medelstora företag*
72. Mohammed Nurul Alam (2002): *Financing of Small and Cottage Industries in Bangladesh by Islamic Banks. An Institutional-Network Approach*
71. Agneta Planander (2002): *Strategiska allianser och förtroendeprocesser – en studie av strategiska samarbeten mellan högteknologiska företag*
70. Anders Bengtsson (2002): *Consumers and Mixed-Brands. On the Polysemy of Brand Meaning*
69. Mikael Hellström (2002): *Resultatenheter i kommunalteknisk verksamhet – struktur, process och effekt*
68. Ralph Meima (2002): *Corporate Environmental Management. Managing (in) a New Practice Area*
67. Torbjörn Tagesson (2002): *Kostnadsredovisning som underlag för benchmarking och prissättning – studier av kommunal va-verksamhet*
66. Claus Baderschneider (2002): *Collaboratively Learning Marketing: How Organizations Jointly Develop and Appropriate Marketing Knowledge*
65. Hans Landström, Jan Mattsson, Helge Helmersson (2001): *Ur en forskar-handledares örtagård. En vänbok till Bertil Gandemo*
64. Johan Anselmsson (2001): *Customer-Perceived Quality and Technology-Based Self-service*

63. Patrick Sweet (2001): *Designing Interactive Value Development. Perspectives and Strategies for High Precision Marketing*
62. Niclas Andrén (2001): *Essays on Corporate Exposure to Macroeconomic Risk*
61. Heléne Tjärnemo (2001): *Eco-Marketing & Eco-Management*
60. Ulf Elg & Ulf Johansson (2000): *Dynamiskt relationsbyggande i Europa. Om hur olika slags relationer samspelar, illustrerat av svenska dagligvaruföretag*
59. Kent Springdal (2000): *Privatisation of the IT Sector in Sweden*
58. Hans Knutsson (2000): *Process-Based Transaction Cost Analysis. A cost management exploration in SCA Packaging*
57. Ola Mattisson (2000): *Kommunala huvudmannastrategier för kostnadspress och utveckling. En studie av kommunal teknik*
56. Karin Bryntse (2000): *Kontraktsstyrning i teori och praktik*
55. Thomas Kalling (1999): *Gaining Competitive Advantage through Information Technology. A Resource-Based Approach to the Creation and Employment of Strategic IT Resources*
54. Matts Kärreman (1999): *Styrelseledamöters mandat – ansats till en teori om styrelsearbete i börsnoterade företag*
53. Katarina Svensson-Kling (1999): *Credit Intelligence in Banks. Managing Credit Relationships with Small Firms*
52. Henrik Kristensen (1999): *En studie av prisförhandlingar vid företags förvärv*
51. Anders H. Adrem (1999): *Essays on Disclosure Practices in Sweden. Causes and Effects*
50. Fredrik Ljungdahl (1999): *Utveckling av miljöredovisning i svenska börsbolag – praxis, begrepp, orsaker*
49. Kristina Henriksson (1999): *The Collective Dynamics of Organizational Learning. On Plurality and Multi-Social Structuring*
48. Stefan Sveningsson (1999): *Strategisk förändring, makt och kunskap. Om disciplinering och motstånd i tidningsföretag*
47. Sten-Åke Carleheden (1999): *Telemonopolens strategier. En studie av telekommunikationsmonopolens strategiska beteende*
46. Anette Risberg (1999): *Ambiguities Thereafter. An interpretive approach to acquisitions*
45. Hans Wessblad (1999): *Omständigheter på ett kärnkraftverk. Organisering av risk och institutionalisering av säkerhet*
44. Alexander Styhre (1998): *The Pleasure of Management Ideas. The discursive formation of Kaizen*

43. Ulla Johansson (1998): *Om ansvar. Ansvarsföreställningar och deras betydelse för den organisatoriska verkligheten*
42. Sven-Arne Nilsson (1998): *Redovisning av Goodwill. Utveckling av metoder i Storbritannien, Tyskland och USA*
41. Johan Ekström (1998): *Foreign Direct Investment by Large Swedish Firms –The Role of Economic Integration and Exchange Rates*
40. Stefan Yard (1997): *Beräkningar av kapitalkostnader – samlade effekter i bestånd särskilt vid byte av metod och avskrivningstid*
39. Fredrik Link (1997): *Diffusion Dynamics and the Pricing of Innovations*
38. Frans Melin (1997): *Varumärket som strategiskt konkurrensmedel. Om konsten att bygga upp starka varumärken*
37. Kristina Eneroth (1997): *Strategi och kompetensdynamik – en studie av Axis Communications*
36. Ulf Ramberg (1997): *Utformning och användning av kommunala verksamhetsmått*
35. Sven-Olof Collin (1997): *Ägande och effektivitet. Wallenberggruppens och Svenska Handelsbanksgruppens struktur, funktion och effektivitet*
34. Mats Urde (1997): *Märkesorientering och märkeskompetens. Utveckling av varumärken som strategiska resurser och skydd mot varumärkesdegeneration*
33. Ola Alexanderson, Per Trossmark (1997): *Konstruktion av förnyelse i organisationer*
32. Kristina Genell (1997): *Transforming management education. A Polish mixture*
31. Kjell Mårtensson (1997): *Företagets agerande i förhållande till naturbelastningen. Hur företaget möter myndigheternas miljökrav*
30. Erling Green (1997): *Kreditbedömning och intuition. Ett tolkningsförslag*
29. Leif Holmberg (1997): *Health-care Processes. A Study of Medical Problem-solving in the Swedish Health-care Organisation*
28. Samuel K. Buame (1996): *Entrepreneurship. A Contextual Perspective. Discourses and Praxis of Entrepreneurial Activities within the Institutional Context of Ghana*
27. Hervé Corvellec (1996): *Stories of Achievement. Narrative Features of Organizational Performance*
26. Kjell Tryggestad (1995): *Teknologistrategier og post Moderne Kapitalisme. Introduksjon av computerbasert produksjonsteknik*
25. Christer Jonsson (1995): *Ledning i folkrörelseorganisationer – den interaktiva ledningslogiken*
24. Lisbeth Svengren (1995): *Industriell design som strategisk resurs. En studie av designprocessens metoder och synsätt som del i företags strategiska utveckling*
23. Jon Aarum Andersen (1994): *Ledelse og effektivitet. Teori og prøving*
22. Sing Keow Hoon-Halbauer (1994): *Management of Sino-Foreign Joint Ventures*

21. Rikard Larsson, Lars Bengtsson, Kristina Eneroth, Allan T. Malm (1993): *Research in Strategic Change*
20. Kristina Artsberg, Anne Loft & Stefan Yard (1993): *Accounting Research in Lund*
19. Gert Paulsson (1993): *Accounting Systems in Transition. A case study in the Swedish health care organization*
18. Lars Bengtsson (1993): *Intern diversifiering som strategisk process*
17. Kristina Artsberg (1992): *Normbildning och redovisningsförändring. Värderingar vid val av mätprinciper inom svensk redovisning*
16. Ulf Elg, Ulf Johansson (1992): *Samspelet mellan struktur och agerande i dagligvarukedjan. En analys ur ett interorganisatoriskt nätverksperspektiv*
15. Claes Svensson (1992): *Strategi i federativa organisationer – teori och fallstudier*
14. Lars Edgren (1991): *Service management inom svensk hälso- och sjukvård – affärsutveckling och kundorganisation*
13. Agneta Karlsson (1991): *Om strategi och legitimitet. En studie av legitimitetsproblematiken i förbindelse med strategisk förändring i organisationer*
12. Anders Hytter (1991): *Den idémässiga dimensionen – decentralisering som struktur och idéförändring*
11. Anders Anell (1991): *Från central planering till lokalt ansvar. Budgeteringens roll i landstingskommunal sjukvård*
10. Rikard Larsson (1990): *Coordination of Action in Mergers and Acquisitions. Interpretive and Systems Approaches towards Synergy*
9. Sven-Olof Collin (1990): *Aktiebolagets kontroll. Ett transaktionskostnads-teoretiskt inlägg i debatten om ägande och kontroll av aktiebolag och storföretag*
8. John Ogbor (1990): *Organizational Change within a Cultural Context. The Interpretation of Cross-Culturally Transferred Organizational Practices*
7. Rikard Larsson (1989): *Organizational Integration of Mergers and Acquisitions. A Case Survey of Realization of Synergy Potentials*
6. Bertil Hultén (1989): *Från distributionskanaler till orkestrerade nätverk. En studie om fabrikanters kanalval och samarbete med återförsäljare i svensk byggmaterialindustri*
5. Olof Arwidi (1989): *Omräkning av utländska dotterföretags redovisning. Metodproblem och konsekvenser för svenska koncerner*
4. Bengt Igelström (1988): *Resursskapande processer vid företagande i kris*
3. Karin Jonnergård (1988): *Federativa processer och administrativ utveckling. En studie av federativa kooperativa organisationer*
2. Lennart Jörberg (1988): *Svenska företagare under industrialismens genombrott 1870–1885*

1. Stefan Yard (1987): *Kalkyllogik och kalkylkrav – samband mellan teori och praktik vid kravställandet på investeringar i företag*

Sizing up leadership

Norms and normativity in and around leadership measures

In a highly quantified world where things tend to only really count if they are counted, this thesis encourages us to critically reflect on the implications and costs associated with putting numbers on phenomena such as leadership, personality, behaviour and reputation.

In my study, I explore the role played by norms and social actors in establishing the acceptability and the purported validity of leadership measures. Taking an interpretivist and critical approach, I uncover the normative agendas and social contexts of four different measurement tools for leadership and personality assessment. I deploy two concepts – normalising potentials and mediating strategies – to argue that we should understand the performative effects of quantitative assessment tools in relation to test practitioners' and test takers' interaction with them.

My study contributes with insights into the link between a measure and its potential performative effects, a link that is often downplayed or assumed automatically established. I argue, that leadership measures rely on social actors' ongoing strategic work, and that any performative effect of the instruments therefore remain a potentiality.

Foregrounding the norms, assumptions, and paradoxes within quantitative measures and the contextual factors their effects rely on, brings to the centre (of discussion and study) the traces of subjectivity and normativity in and around the measures. These insights allow us to better challenge and question the ascribed authority and mandate of quantitative assessment tools. Doing so, we might reduce the dichotomy between perceived objective, trustworthy quantitative tools and subjective opinions and beliefs, ultimately making room for alternative understandings of and approaches to leadership (development).

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

SCHOOL OF ECONOMICS AND MANAGEMENT

Department of Business Administration

ISBN 978-91-8039-551-9