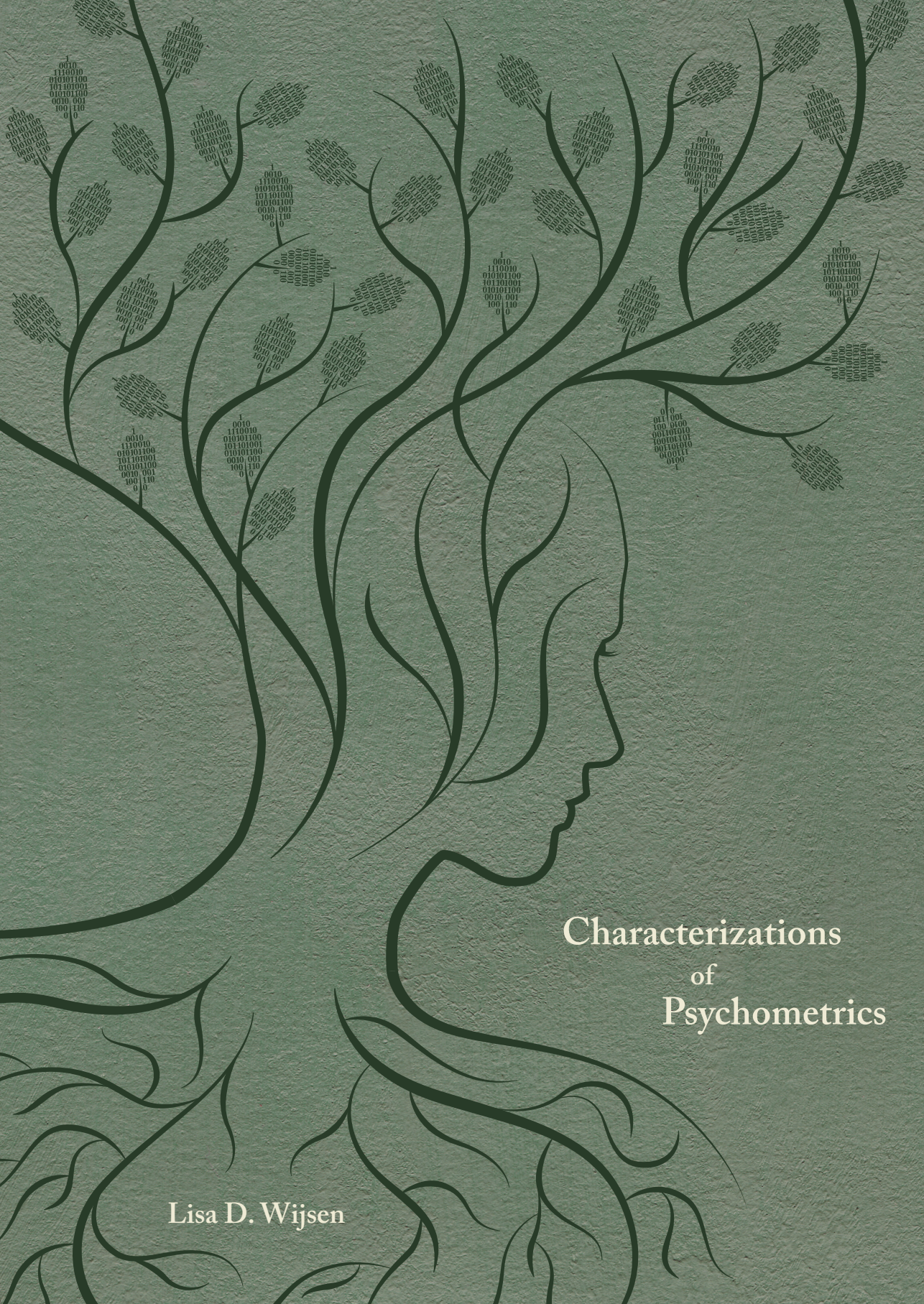
Characterizations of psychometrics

Wijsen, L.D.

Publication date
2021
Document Version
Final published version



Characterizations of Psychometrics

Lisa D. Wijsen

Characterizations of Psychometrics

Lisa D. Wijsen

2020

Characterizations of Psychometrics

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex

ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen
op 22 maart 2021

door

Lisa Dide Wijsen
geboren te Utrecht

The research in this dissertation was funded by a NWO Graduate Programme Grant (given out by IOPS).

ISBN: 978-94-92332-33-2
Cover & Layout: Esther Scheide, proefschriftomslag.nl
Printed by: ProefschriftMaken

Promotores:	Prof. Dr. D. Borsboom	Universiteit van Amsterdam
	Prof. Dr. W. J. Heiser	Universiteit Leiden
Overige leden:	Prof. Dr. Agneta Fischer	Universiteit van Amsterdam
	Prof. Dr. Han van der Maas	Universiteit van Amsterdam
	Prof. Dr. Rens Bod	Universiteit van Amsterdam
	Dr. Anna Alexandrova	Cambridge University
	Dr. Femke Truijens	Erasmus Universiteit
	Prof. Dr. Klaas Sijtsma	Tilburg Universiteit

Contents

Chapter 1

General Introduction

Psychometrics is best-known for its applications that so many of us are familiar with. When Dutch school children are about 11 or 12 years old – and I remember this clearly –, they participate in what is called the CITO test: an elaborate assessment that provides information on a child's progress on skills like reading ability and mathematics, and helps guide teachers in giving advice on a child's future school choice. When we grow older and apply for a job, we may be asked to do an assessment that maps the applicant's desirable skills and traits, and for a person with mental health problems, measurement instruments like the Beck Depression Inventory or the Hamilton Anxiety Rating Scale might help in determining the appropriate diagnosis. In different phases of our lives, we encounter psychometric applications, which can be very influential or even decisive in high-stake situations. Was your job assessment satisfactory to your future employer? Are your SAT scores high enough for admission into a top-notch Ivy-League university? However, even though most of us are familiar with different examples of psychometric applications, psychometrics as a research area is much more mysterious.

Psychometrics is a very complex scientific discipline: its content is often highly technical, heavily embedded in statistics, and its connection with psychological or educational applications is often difficult to grasp (and sometimes, on the face of it, non-existent). These aspects make psychometrics' purposes, goals, values, and concepts hard to understand for an outsider who is not part of the psychometric community. So, what is it that psychometricians actually do, and why do they do it? In this dissertation, I explore different aspects of the scientific domain of psychometrics – its disciplinary structure, its historical development, its practices, its models, and its values – to improve our understanding of this lesser-known and complex discipline.

Psychometrics as understood in this dissertation is the field that concerns itself with the measurement and prediction of human behavior. Psychometric research usually entails a thorough investigation into the development and/or application of one or more psychometric models, which are often members of the families of Item Response Theory (IRT, Embretson & Reise, 2000; Van der Linden & Hambleton, 2013), Structural Equation Modeling (SEM, Jöreskog, 1970; 1973), Multidimensional Scaling (Kruskal & Wish, 1978), and, particularly at my university, Network Models (Borsboom & Cramer, 2013; Epskamp, Rhemtulla, & Borsboom, 2017). IRT models are tools for analyzing testing data (item responses), by modeling the relationship between item responses and a latent variable (an unobservable characteristic or attribute such as reading ability or intelligence). Structural Equation Modeling is a method for analyzing structural analyses among two or more latent variables, for example, the structural relationship between reading ability and spatial reasoning.

Originally, psychometric research mainly concerned itself with methods for the analysis of test data, and developed concepts such as reliability, validity, and factor

analysis (possibly three of the most influential and most exported psychometric concepts). Educational measurement, the domain of IRT, is still one of the most important focus points of psychometrics, but throughout the 20th century, psychometric models (and especially factor and SEM models) have been applied to quantitative research in a wide variety of domains such as personality (McCrae & Costa, 1987; Marsh et al., 2010), mental health (Compas et al., 2006; Caspi et al., 2014), and marketing research (Baumgartner & Homburg, 1996). Moreover, not all psychometric models are strictly measurement models. An important branch in psychometrics that came up in the 1960s is Multidimensional Scaling (MDS), which is – very generally put – a method for the representation of similarities and dissimilarities in data and is a popular method in consumer research. All in all, applications of psychometric methods are galore, both in psychology and in other fields.

1.1 Positioning Psychometrics

So, where exactly is the connection between the highly technical psychometrics as described above, and the familiar applications such as the CITO test, mental health diagnostics, or job assessments? The psychometricians whose research and practice are central to this dissertation are often responsible for the analysis of large data sets (such as the data sets resulting from the CITO test or any other large assessment), and develop reliable tools and methods that enable psychometric analysis. Several of them are members of the Psychometric Society, one of the most important international institutions for psychometric research. Importantly, they are not the people who actually write test items or administer tests[1]. They are what we call β-psychometricians (Wicherts, 2007). β-psychometricians have a technical and statistical mind and aim to find optimal ways of analyzing data. Often, they do not have a specific substantive interest, in the way a psychologist does who is interested in depression among adults, or a sociologist who is interested in the anti-vaccination movement in the Netherlands. Contemporary psychometricians often aim to find solutions for the analytic problems that occur in applied research, like psychology or educational measurement, but it is not the substantive part of research that makes them tick: it is the statistical, the abstract, and the technical that are main motivators for psychometricians.

Psychometrics and its substantive counterparts, psychology and educational measurement, have different research agendas and over time, have grown apart (Borsboom,

[1] The conceptualization of psychometrics in this dissertation is relatively narrow, and there is no rule that states we should not include test makers or behavioral data analysts as psychometricians. In the discussion, we elaborate on the definition of psychometrics we used and how it could possibly be expanded in future research.

2006; Sijtsma, 2006; Groenen & Van der Ark, 2006). This detachment between the substantive and the technological has not always been as prominent as it is now: early psychometricians were often psychologists who were heavily engaged in a substantive problem (e.g. the nature of individual differences) and simultaneously tried to solve or explain this phenomenon through technically advanced methods. Nowadays, the shared ancestry with psychology has become somewhat invisible. Psychometrics is incredibly technical, quantitative, and heavily embedded in statistics; it is highly specialized and difficult to understand for anyone without a thorough training in statistics or psychometrics. The relationship between psychometric research and topics in psychology or educational measurement is hard to grasp. It is therefore hard to escape the feeling that psychometrics has become an independent discipline that is sometimes more embedded in statistics than in psychology.

Psychometrics has thus created the impression of operating as an ivory tower: a scientific domain that mainly works on its own problems and has difficulty connecting to other parties. What is incredibly intriguing to me and what plays a role in each of the following chapters, is the constant balance between the content-neutral and the statistical on the one hand, and psychometrics' aim to improve and consult on psychological and educational research and practice on the other hand. The abstractness of psychometrics invites the impression that it is almost mathematics-like in its approach and ideals: whether or not it finds an application seems less important than the technical or mathematical 'beauty' or complexity of the work. In extremis, this means that the more technical the research, the better. However, this is simply not true: historically, psychometricians often had very social or political motives – a mission to improve society or quantitative applied research -, and as we will see in Chapter 3 and Chapter 6 as well, for many contemporary psychometricians, application and impact of their work are still crucial. And that is where the tension comes in: psychometricians often want to have an impact on applied research and the testing industry, but it is very difficult for outsiders, and that includes me, to see how and where they want to make contributions since their work is so technical. Unpacking psychometrics, the purpose of this dissertation, will aid in understanding the identity, goals, and values of this somewhat curious but influential discipline.

A dissertation on the dealings of psychometrics comes at an interesting and perhaps even urgent time. Psychometrics has known a very successful 20th century, when standardized tests and assessment procedures were considered an important stepping stone to providing solid and fair education for everyone, and the development of psychometric models and tools was a fruitful enterprise. However, public support for standardized testing and psychometrics as we know it – specialized in the ranking of people based on ability – has slowly dwindled, and has now come to a new phase. In May 2020, the University of California has voted to stop requiring an SAT or ACT score (both

large standardized tests used for college admission) for freshman applicants. The SAT and ACT tests are considered unfair with regard to poor, Hispanic, or Afro-American students and supposedly counteract diversity and inclusivity. The University of California, encompassing 10 schools including UCLA and UC Berkeley, aims at building its own standardized test that does not show the same bias. If developing an unbiased test were to fail by 2025, the University of California would still decide against using SAT or ACT scores for student admission. It is expected that more American universities will adopt the same policy. Though standardized testing has always received a fair amount of criticism often in the shape of a similar accusation with regard to racial bias, the rejection of SAT and ACT by a major university is quite a turn of events, and it puts psychometrics in a complicated and perhaps fragile position. The underlying ideology of psychometrics, that of the test being able to objectively select the 'best people', is actively being disputed, and the credibility of psychometrics is being questioned. A dissertation on the identity, goals, and values of psychometrics therefore, accidentally, proves quite timely. It is time for psychometrics to pause and take a more reflective stance on its own position as a scientific discipline. Hopefully, this dissertation can make a contribution to that reflective stance.

1.2 Different Perspectives on Psychometrics

A classic psychometric dissertation often entails a thorough investigation into the functioning of one or more psychometric models, such as the ones mentioned above. Such a dissertation could include chapters on possible extensions of already existing models, applications to empirical data, simulation studies, analysis of goodness-of-fit measures, or new methods for parameter estimation. However, this dissertation is not psychometric in its approach: instead, psychometrics is the object of investigation. This dissertation is *about* psychometrics. Psychometrics as a field remains largely under the radar; it is relatively invisible to outsiders, while its output is incredibly influential and has contributed to many features of modern-day society (e.g. the rise of a national education, the prominence of testing in so many parts of our lives, the appreciation of certain cognitive skills like intelligence). My investigation is a start in making psychometrics accessible to people who are not part of the psychometric community, but are curious about what this field actually does. I unpack psychometrics by taking on different roles or perspectives: that of the historian, that of the ethnographer, and that of the philosopher of science. Below, I will elaborate on these different roles and explain how they relate to each of the chapters.

1.2.1 The Historian

The original plan for my dissertation was in fact to write a history of psychometrics. However, this project turned out much broader and I often drifted away from our historical intentions, simply because other perspectives on psychometrics were equally

appealing. Nevertheless, the history of psychometrics still makes up a significant part of this dissertation, and is most prominently present in Chapter 2, in which we constructed an academic genealogy of Psychometric Society presidents. In this genealogy, we have traced back the academic lineages of prominent psychometricians. The genealogy shows that the presidents can be divided in roughly two groups: one group, the majority, descends from psychologists like Wilhelm Wundt and William James, the other smaller group descends from mathematicians and are part of the Carl Gauss lineage. The genealogies show how psychometrics has historically developed, but also how the identity of psychometrics is indeed made up of different branches: the larger branch of psychology and educational measurement, and the smaller branch of mathematics.

The historical approach also comes back in Chapters 3 and 6. In Chapter 3, we asked psychometricians to reflect on the history of psychometrics. For example, what do they consider the most important contribution in the history of psychometrics? And whose work were they really inspired by? In Chapter 6, we draw a comparison between the use of values in psychometric research in the early days of psychometrics, when psychometrics was quite explicit about its political and social motivations, and contemporary psychometrics, in which moral values are more implicit. The purpose of both Chapters 3 and 6 will be explained in more detail below.

1.2.2 The Ethnographer

An ethnographer investigates groups of people or cultures that share certain habits and rituals. It is not a stretch to consider researchers as such a specific group, and from the 1970s onwards, several ethnographies of (laboratory) sciences have been conducted by social scientists and proponents of Science & Technology Studies (e.g. Barnes, 1977; Latour & Woolgar, 1986; Latour, 1987). In Chapter 3, we followed the ethnographic approach and treated psychometricians as a group of people with their own culture that we wanted to uncover. Inspired by oral history methodology (Abrams, 2010; Thompson, 2017), we interviewed twenty Psychometric Society presidents to investigate how prominent psychometricians perceive their own field. In semi-structured interviews, the presidents shed light on their own motivation to become a psychometrician, on the complicated position of psychometrics among other disciplines, on what they consider important historical achievements of psychometrics, and on possible future directions of the discipline. In this chapter, we analyzed how the psychometricians themselves conceptualize psychometrics and we show that there is not one approach to practicing psychometrics. What motivates or inspires psychometricians turns out to be a variety of things: theory construction, building technologically advanced assessment procedures or tests, providing applied researchers with methodological advice, or proving mathematical theorems. This chapter holds a qualitative analysis of these interviews, and I use a

selection of quotes from the interviews to illustrate the main themes. Edited versions of the full interviews will be published as a stand-alone book (Wijsen, forthcoming) that can be read as a complement to the present thesis.

1.2.3 The Philosopher of Science

In Chapters 4, 5, and 6, I take on the role of the philosopher of science, and address questions on the meaning of psychometric models and values in psychometric research. Chapters 4 and 5 zoom in on what it means to give a causal reading of psychometric models and why such a reading can be useful and appropriate. The modeling tradition in psychometrics started with Spearman's (1904) common factor model. In this model, the latent variable denotes general intelligence, which presumably had a causal effect on item scores: because of a person's level of general intelligence, he or she scores higher on cognitive tests than people with a lower level of general intelligence. Chapter 4 provides an overview of three different causal modeling traditions: reflective models (an example of which is Spearman's common factor model), formative models, and network models. In Chapter 5, we argue why a causal reading of the common factor model can be appropriate and desirable. Namely, when there is sufficient reason to believe that the underlying causal structure of the data is indeed a common causal structure, a causal reading of this model offers several benefits that a statistical reading does not.

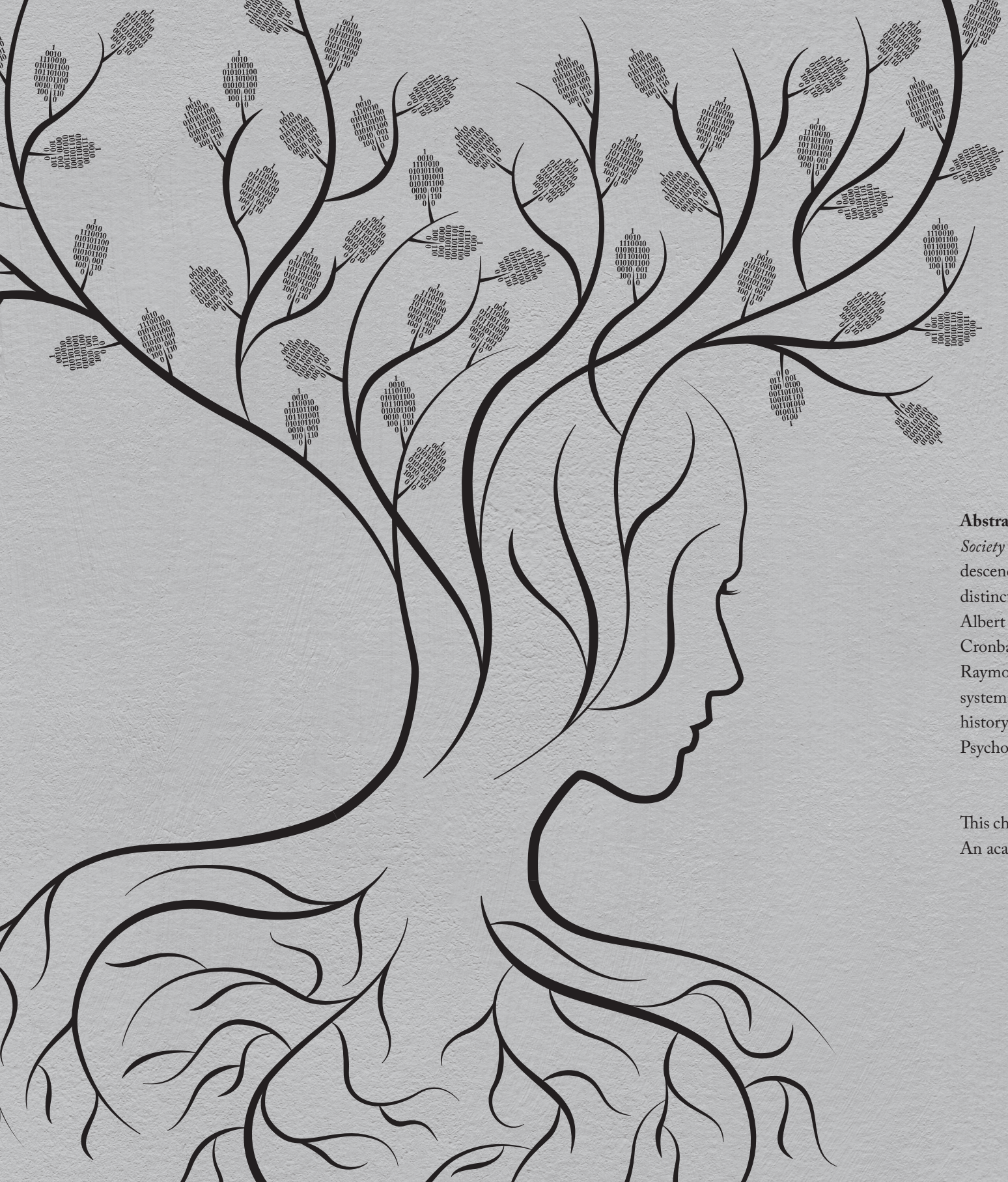
Note that Chapters 4 and 5 are not descriptions of practice in psychometric research per se: they offer a possible reading for psychometric models that we think is beneficial in specific circumstances, namely, when the modeler is interested in building substantive theory and aims at formulating explanations for psychological phenomena. Interestingly enough, even though a causal interpretation has several benefits over a statistical interpretation, using psychometric models as causal models these days is a rarity. Latent variables or common causes are often not believed to be entities, but are considered practical in the modelling process. And perhaps unsurprisingly, these models are not often used for theory building. In fact, in Chapter 6, we find that in psychometrics, it is utility rather than truth, that is considered a value. In the General Discussion, I will elaborate on how a different reading of psychometric models, namely a pragmatist interpretation, could improve our understanding of psychometric practice.

Chapter 6 is an analysis of underlying values in contemporary psychometrics. Science is often expected to be practiced without any involvement of personal preferences or social values, i.e. the value-free ideal. Research is a field that prides itself in having its own unique culture that is drastically different from other domains, such as religion or politics, because of its focus on a set of norms or epistemic values, like impartiality, objectivity, and transparency, which lead to reliable knowledge production (Mulkay, 1976). Contemporary psychometrics invites the image of being mainly an abstract, technical and

socially uninvolved discipline. Its research questions are often formulated as technical or statistical problems that require an equally technical or statistical solution (Evans & Waites, 1981), inviting the impression that it is a value-free discipline indeed. However, as philosophers of science have shown over the past decades, all scientific disciplines, even those of a technological and statistical nature, incorporate social values in several aspects of the research process (Douglas, 2000; 2009; Elliott, 2017), and so does psychometrics. In this chapter, we analyze the role of values in contemporary psychometrics. The values we discuss are the conceptualization of individual differences as quantitative (not qualitative) differences, the aim for objective measurement, the aim for fair measurement, and the preference for utility above truth.

1.3 The Merit of this Dissertation

I want to stress here that it is by no means my intention to create the impression that I have done a full ethnography, a full history or a full philosophy of psychometrics – each of these approaches alone could have resulted in one or more dissertations –, or that I have now acquired the skills and knowledge that each of these professions require. Instead, I have borrowed parts of each approach to shed light on the interesting object that is psychometrics. Even though I have taken up a number of different roles and perspectives in this project rather than use one approach throughout – which may seem eclectic at first – there is merit in this, and this is a different type of merit than taking on a singular, more detailed approach. What I aim to do is investigate different parts of the same research object and thereby draw different characterizations of psychometrics. As such, it opens up a discipline that seems incomprehensible and mysterious to people who are not part of its community. It gives those who are curious about psychometrics an idea of its historical identity, values, and practice. By dividing my attention over a number of topics and approaches, this dissertation has created a variety of new starting points for further research. I consider this work an open invitation to philosophers, historians, sociologists of science, and fellow psychometricians to take up one of these starting points and investigate further what it is that psychometricians actually do, how this has come to be, and what psychometricians ought to be doing.



Chapter 2

An Academic Genealogy of Psychometric Society Presidents

Abstract In this paper, we present the academic genealogy of presidents of the *Psychometric Society* by constructing a genealogical tree, in which Ph.D. students are encoded as descendants of their advisors. Results show that most of the presidents belong to five distinct lineages that can be traced to Wilhelm Wundt, James Angell, William James, Albert Michotte or Carl Friedrich Gauss. Some important psychometricians like Lee Cronbach and Charles Spearman play only a marginal role, others (Francis Galton, Raymond B. Cattell, and Gustav Fechner) are even completely absent. The genealogy systematizes important historical knowledge that can be used to inform studies on the history of psychometrics, and exposes the rich and multidisciplinary background of the Psychometric Society.

This chapter is adapted from Wijsen, L. D., Borsboom, D., Cabaço, T., Heiser, W. J. (2019). An academic genealogy of Psychometric Society presidents. *Psychometrika, 84*, 562 - 588.

2.1 Introduction

Psychometrics is a scientific discipline concerned with "quantitative measurement practices in psychology, education and the social sciences" ("What is psychometrics?", n.d.). The origin of psychometrics is often traced to the primary example of a model for measurement: the common factor model, constructed by Charles Spearman (1904), in which a set of observed variables is regressed on a common latent variable. This model provided the classical decomposition of test scores into general and item-specific sources of variance that enabled psychologists and psychometricians to systematically measure what Spearman called *general intelligence,* or *g*. Psychometric research, commonly understood to cover the technical rather than the substantive side of test theory (Borsboom, 2006), has further developed and branched out in a wide array of modeling techniques such as Classical Test Theory (CTT), Structural Equation Modeling (SEM), Item Response Theory (IRT), and Multidimensional Scaling (MDS) (Jones & Thissen, 2007).

Variations of these models have been used for the measurement and prediction of a multitude of psychological attributes, such as personality dimensions (McCrae & Costa, 1987), mental abilities (Thurstone, 1938; Carroll, 1993), and psychiatric disorders (Caspi et al., 2014). In addition, implementations of psychometric testing have resulted in a massive set of practical test applications; examples include college admission tests (e.g. the SATs or the Medical College Admission Test), various psychological assessments for job performance (Bowling & Hammond, 2008), and clinical assessments used in psychiatric practice (Floyd & Widaman, 1995). As a result, at the beginning of the 21st century, circumventing the psychological test has become almost impossible in large parts of the world.

Even though psychometrics has thus taken up a prominent position in the scientific domain, and in society at large, little research has been done on its origins and its development through time (exceptions are: Dehue, 1995; Groenen & van der Ark, 2006; Jones & Thissen, 2007; Van der Heijden & Sijtsma, 1996). As a result, we have a limited view of the lines of intellectual descent that have led to the current structure of the psychometric field. This article aims to contribute to a better understanding of these issues by providing an academic genealogy of Psychometric Society presidents. An academic genealogy is a genealogical tree, in which the traditional ancestor-descendent relations are replaced by relations between doctoral advisors and their students (Kealy & Mullen, 1996). As such, the academic genealogy provides an overview of one or more scientific disciplines through a simple but clear visualization of advisor-student relations.

In historical research, academic genealogies have been used to trace the descent of one single person (Bennet & Lowe, 2005; Williams, 1993), to identify causes of unresolved disagreements in medical science (Hirshman et al., 2016), or for tracing scientific ideologies back to their roots (Robertson, 1994). The use of genealogical trees has also proven popular

outside professional historical research, as can be seen from the fact that several websites have been developed for this purpose (e.g., *PhDTree, the Mathematics Genealogy Project*, *Neurotree*). Finally, throughout the sciences there is a general trend of constructing academic genealogies to uncover the evolution of individual disciplines, such as mathematics (Gargiulo, Caen, Lambiotte, & Carletti, 2016), astronomy (Tenn, 2016), biology (Bennett & Lowe, 2005), psychology (Williams, 1993; Robertson, 1994) and primatology (Kelley & Sussman, 2007). In the current study, we aim to contribute to this development by constructing a well-documented and verifiable academic genealogy of Psychometric Society presidents, that can function as a resource for historians of psychometrics and as such may help to yield insights into the questions of how psychometrics originated as a new scientific discipline and how it has developed over time.

The structure of this paper is as follows. The methods section will elaborate on the methods used for the collection of the data and the construction of the genealogies. In the results section, five distinct genealogies are presented, each accounting for a number of presidents. In the discussion section, we will provide some ideas for further historical research, and discuss some limitations.

2.2 Methods

To minimize selection bias while keeping the practical task within reasonable bounds, we have chosen to construct a genealogy in which presidents of the Psychometric Society function as the initial set, from which we work 'backwards in time' to uncover each of their lineages. There are several reasons for selecting the presidents of the Psychometric Society as our starting point. The Psychometric Society, originally founded in 1935 by L.L. Thurstone and others, is the most important institution concerned with psychometric science, and *Psychometrika* (founded by Paul Horst in 1936) its flagship journal (Heiser & Hubert, 2016). By organizing the most central meeting of psychometricians – the International Meeting of the Psychometric Society (IMPS) – the Psychometric Society is the main social structure in the psychometric community. As a result, presidents of the Psychometric Society, elected by the members of the Psychometric Society, are invariably central figures in the discipline, with important social and scientific functions in psychometrics. Therefore their genealogy is both intrinsically interesting and likely to contain important historical information. It goes without saying that when we state that presidents of the Psychometric Society are typically important psychometricians, it does not imply that psychometricians who did not become president are in any way less important or anything of the sort. Hence, the genealogy covers only the history of the Psychometric Society. However, given the requirement of the current study, that each link in the tree must be backed by historical evidence, practical constraints necessitate that we leave extending the tree beyond this set to a future study.

Up until 2017, the Psychometric Society has counted 84 presidents, including the president-elect, all of whom have taken up prominent positions in the field of psychometric research in their time. By taking these 84 presidents as our initial set of nodes, the genealogy was constructed in an objective, systematic fashion: each individual advisor-student relation has been assessed through careful archival research and is adequately documented. However, it is important to note that, because the initial set in the tree represents the presidents of the Psychometric Society, other nodes include *only* those scholars that are necessary for connecting these presidents back to their ancestors. This means that nodes were added *only* for (a) representing a president directly, or (b) tracing the line of descent of a president back through time. Once different lineages shared a common ancestor in the 19th century, further investigation of the line of descent was stopped. This both served to limit the number of nodes added to the tree and to ensure that the genealogical tree would not stretch out over too long a period of time. When an individual lineage did not share any common ancestors with any other lineage, we aimed to trace the lineage back into the 19th century.

To collect all the necessary information on student-advisor relations, several archival sources were consulted. These included doctoral dissertations, obituaries, monographs, resumes, university websites and the *Mathematics Genealogy Project*. Online academic genealogies that are open for edits by users, such as *Neurotree* (David & Hayden, 2012) or *PhDTree* (now taken offline) proved to be helpful sources as well, though connections taken from these websites were always double checked for accuracy. When online resources were not conclusive, we also engaged in personal communication with presidents or with archivists of university libraries. In the data collection phase, we coded each person's supervisor, their year of dissertation and the university where they wrote their dissertation.

To ensure a systematic approach we applied a number of rules. Firstly, we have only included relations that are of the advisor-graduate student kind, and thus excluded any other influential relations (e.g. friendships, informal mentoring, cooperation among colleagues). Secondly, we have only included one advisor per person. Thirdly, we only included advisors of people who actually finished their dissertation and got an official Ph.D. degree. If people do not have a Ph.D. degree, they are simply the start of their lineage.

2.3 Results

Following the rules we set in the methodology section, the total academic genealogy of psychometrics includes 208 scholars, of which 84 are presidents. This genealogy comprises 5 separate genealogies, leading back to James R. Angell, Wilhelm Wundt, William James, Carl Friedrich Gauss or Albert Michotte. 64 out of 84 presidents connect to one of these genealogies.

Sixty-six out of 84 of the presidents received their Ph.D. degree in the United States, 14 in Europe, 1 in Asia, 1 in Africa and 1 in Australia. Five presidents are female, who are also the only women in this genealogy. Universities that have produced a large number of presidents are University of Chicago (n = 14), Princeton University (n = 9), Columbia University (n = 8), the University of Illinois (n = 6) and Stanford University (n = 5)

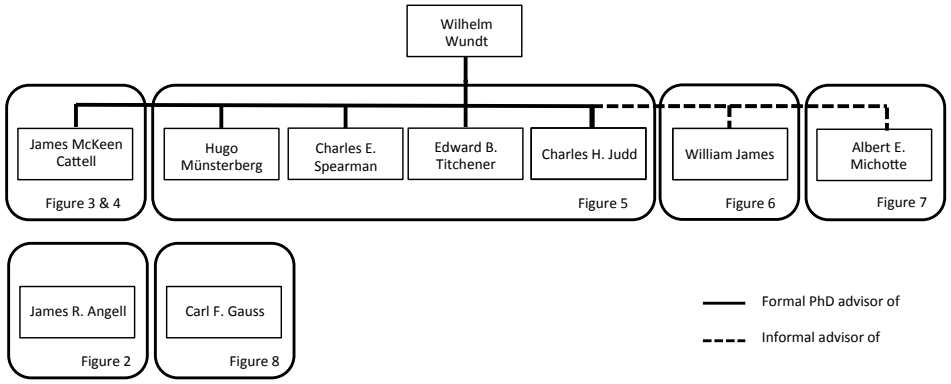


Figure 1. Overview of the different genealogies in this paper. The genealogies of Cattell, Münsterberg, Spearman, Titchener and Judd, doctoral students of Wundt, will be discussed in Figures 3, 4 and 5. Figures 6 and 7 show the genealogies of James and Michotte (both non-doctoral students of Wundt), and the genealogies of Angell and Gauss in Figure 2 and Figure 8.

Figure 1 provides an overview of the 5 different genealogies, which are visualized in Figures 2 to 8. Figures 3, 4 and 5 show the genealogies of the official doctoral students of Wilhelm Wundt. Due to James McKeen Cattell's sizeable lineage, it has been divided over two figures. The genealogies of James R. Angell (Figure 2), William James (Figure 6), Albert E. Michotte (Figure 7) and Carl Friedrich Gauss (Figure 8) have no official connections[2] to Wilhelm Wundt, and are given separately. Table 1 provides the lineages of the presidents that are not part of any of the larger genealogies. Tables 2 to 6 (see Appendix A) provide the references of the sources we used for each advisor-student relation in this genealogy.[3]

[2] An official connection here is the relationship between doctoral students and their advisors. James, Michotte, Angel, and Gauss were not Wundt's official doctoral students, though some knew Wundt quite well.

[3] As the source for the Gauss genealogy is the Mathematics Genealogy Project, we have excluded a reference table from the Supplementary Material.

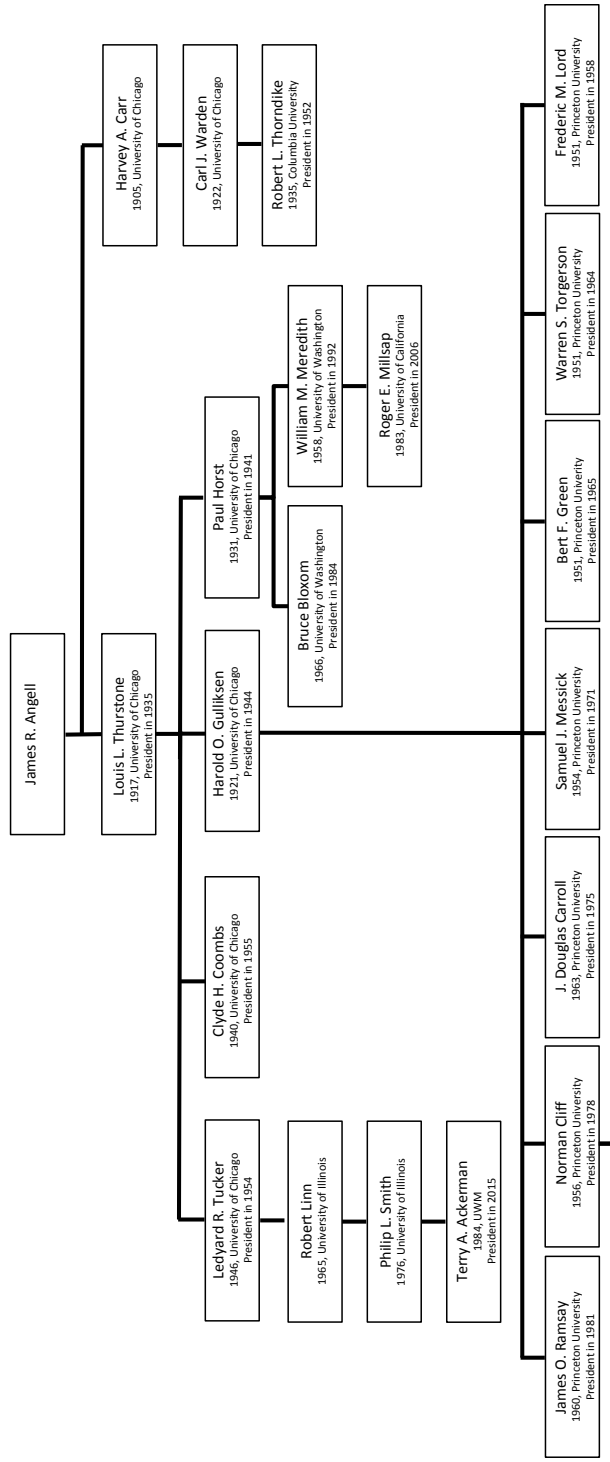


Figure 2. The genealogy of James R. Angell. For each scholar in the genealogy, the year of obtaining a Ph.D. Degree and the university in question, are given. For presidents of the Psychometric Society, the year of their presidency is included.

2.3.1 The Genealogy of James R. Angell

The first genealogy we discuss starts with James R. Angell, an American psychologist who never officially finished his dissertation, and is hence the starting point of this tree (see Figure 2). Angell's genealogy holds 25 scholars of which 20 are presidents of the Psychometric Society. One of Angell's major accomplishments was his presidency of Yale University from 1921 until 1937. One of Angell's students was Louis L. Thurstone, the founder and first president of the Psychometric Society. Thurstone became very well known for his work on scaling (Thurstone, 1927) and multiple factor analysis (Thurstone, 1934; 1935). Four of the psychometricians who descended from Thurstone were president of the Psychometric Society in its early days: Ledyard R Tucker, Clyde Coombs, Harold Gulliksen, and Paul Horst.

Paul Horst, president in 1941 and 1942, was one of the six founders of *Psychometrika*, and successfully pursued the idea of setting up a journal dedicated to mathematically oriented psychological research (Heiser & Hubert, 2016). Many others supported him in this endeavor, among whom future presidents Jack Dunlap, Joy P. Guilford (see Figure 5) and Harold Gulliksen (see Figure 2). Years later, two of Horst's students became presidents of the Psychometric Society: William Meredith, in 1992, and Bruce Bloxom, in 1984 (see Figure 2). Meredith was the advisor of Roger E. Millsap, who shared with him an interest in measurement invariance, and who became president in 2006.

Ledyard R Tucker, president in 1954, wrote his dissertation in 1941. In his early career, he worked on factor analysis together with Thurstone. From 1944 until 1960, he became the first director of statistical analysis at the Educational Testing Service (ETS) and worked as a lecturer at Princeton University. After that he became professor of psychology at University of Illinois. As Bert F. Green (1980) eloquently puts it, Ledyard Tucker had a lifelong "affair with psychometrics". Tucker's lineage goes to Robert Linn, educational psychologist, via Philip L. Smith, to Terry Ackerman, who was president in 2015. Ackerman now holds the "Lindquist Chair" at ACT.

Harold Gulliksen, president in 1944, turns out to be one of the most fruitful suppliers of presidents of the Psychometric Society in the entire genealogy: seven presidents were directly advised by Gulliksen, and another three are of the next two generations, advised by either Norman Cliff or Forrest W. Young (see Figure 2). During the Second World War, Gulliksen worked for the Navy, where he introduced many different selection and assessment instruments. He noted that assessment instruments should not only include tasks that reflect skills they learn during the training, but also skills they actually employ on the job (Messick, 1998). Gulliksen's *Theory of Mental Tests* (1950) was one of the first comprehensive handbooks on classical test theory. In 1945 he was appointed professor of psychology at Princeton University until 1972 and also worked at ETS after WWII as a research advisor until 1974, doing both half time. Besides an interest in mental testing,

Gulliksen also had a strong interest in learning and used mathematical models to visualize learning curves (Messick, 1998).

Another president and student of Thurstone who was well known for his interest in mathematical psychology was Clyde Coombs, president in 1955. Throughout his career, he published on analysis of qualitative structures of proximity and dominance data (Coombs, 1952; 1964) and models of conflict and choice (Tversky, 1992).

A surprising presence in Angell's genealogy is Robert L. Thorndike, the son of the famous Edward L. Thorndike (who is part of Wundt's genealogy (Figure 3), not Angell's). After writing a dissertation on animal learning, he specialized in educational psychology and made contributions to the field of reliability. Like his father, he stayed at Columbia University throughout his entire career and worked closely together with Irving Lorge (Figure 3). He was president in 1952.

Harold Gulliksen's graduate students.

As mentioned earlier, Gulliksen was extremely productive when it came to preparing his students for the presidency of the Psychometric Society: no less than seven of his students eventually became president of the Psychometric Society. The first three to write their dissertations, in 1951, were Warren S. Torgerson, Bert F. Green, and Frederic M. Lord.

In his dissertation, Torgerson developed a method of multidimensional scaling (MDS), a new research area that gained a prominent presence in psychometric research. After finishing his degree, he spent some time working for the Navy, where he headed the statistical analysis division. In 1955, he returned to Princeton University and resumed his work on MDS (Green, 1999). Bert F. Green was president in 1965, one year after Torgerson's presidency in 1964, and was editor of *Psychometrika* between 1972 and 1980.

Frederic M. Lord, president in 1958, was one of the most famous psychometricians in the history of psychometrics. Together with Melvin Novick (see Figure 8), he wrote what is perhaps the most significant textbook in the history of psychometrics: *Statistical Theories of Mental Test Scores* (1968). In this book, Lord and Novick provided a mathematical account of classical test theory and an introduction into modern test theory. From 1949 onwards, Lord worked at ETS as Director of Statistical Analysis, and wrote his dissertation at Princeton in 1951. During this time, he became an important proponent of Item Response Theory (IRT).

J. Douglas Carroll wrote his dissertation in 1963, and left for Bell Labs as Member of Technical Staff after finishing his dissertation (Heiser, 2012). From 1925 until 1996, Bell Labs was the research center of AT&T, where many scientists of a wide variety of disciplines worked side by side on new ideas for information technology. Carroll's dissertation still focused on human learning, but due to the influence of many fellow psychometricians who also worked at Bell Labs (among others Roger Shepard, Figure 6,

and Joe Kruskal, Figure 8) he made the switch to MDS. Over the years, Carroll invited many other psychometricians to spend some time at Bell Labs, and Bell Labs became the most important center for MDS-related research.

Samuel Messick wrote his dissertation in 1954, after which he joined ETS in 1956, where he stayed until he retired as vice-president in 1997. One of his main interests was the concept of validity, in which he stressed the idea that validity was not purely a property of a test, but also a property of the interpretation the test scores (Messick, 1989b). Norman Cliff wrote his dissertation in 1957, succeeded Bert Green as editor of *Psychometrika* (1981-1984), and supervised two presidents of the Psychometric Society: Robert Cudeck, profesor in quantitative psychology at Ohio State University, and Forrest W. Young. Forrest Young was professor in quantitative psychology at the University of North Carolina, and was the advisor of Yoshio Takane, president in 1986. James O. Ramsay, famous for his work on functional data analysis, was the last president of the Psychometric Society to write his dissertation under Gulliksen's supervision.

2.3.2 The Genealogy of Wilhelm Wundt, part I

The largest genealogy starts with Wilhelm Wundt. Wundt's genealogy counts 42 scholars, of whom 23 are presidents. Due to its scope, the entire Wundt genealogy is divided over three graphics: Figure 3, Figure 4 and Figure 5.

The German Wilhelm Wundt is often seen as the founder of experimental psychology, and led the first psychology laboratory in Leipzig. Wundt oversaw an exceptional number of 184 doctoral dissertations between 1875 and 1919, and many became famous psychologists or psychometricians, five of whom can be found in this genealogy: James McKeen Cattell (Figure 3 and 4), Hugo Münsterberg, Charles Spearman, Charles H. Judd and Edward Titchener (Figure 5). Wundt's influence was not limited to Europe; he also attracted many American psychologists, like Cattell and Titchener, who travelled all the way to Europe to spend time at Wundt's laboratory. Inspired by physiological experiments, Wundt initiated a line of research in which psychological objects, such as sensory stimuli, were investigated through experiments (Danziger, 1994). Wundt believed that experiments were only suitable for direct responses to physical stimuli that are bereft of any interpretation. The participants in his experiments were his own students, and they were thoroughly trained in how to accurately report their responses. Important to note here, is that Wundt's psychology was not limited to experiments. He found experimental psychology the most suitable method for discovering psychological laws, but he believed that psychological experiments should not take up the entire space of psychological science. He considered topics like language, myth, culture and religion not suitable for experimental psychology, and used his *Völkerpsychologie* as a method for investigating those (Danziger, 1994).

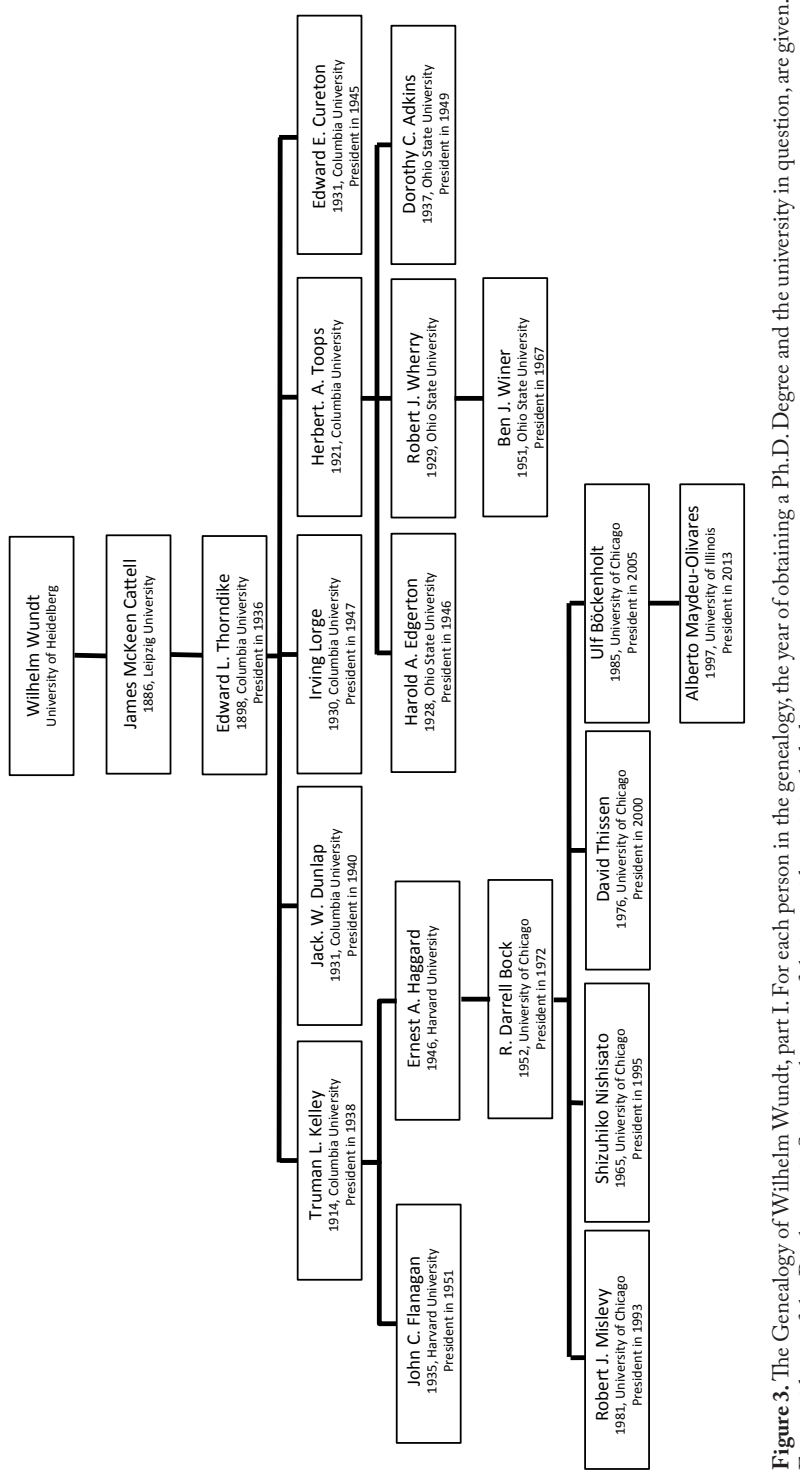


Figure 3. The Genealogy of Wilhelm Wundt, part I. For each person in the genealogy, the year of obtaining a Ph.D. Degree and the university in question, are given. For presidents of the Psychometric Society, the year of their presidency is included.

Though the five direct descendants of Wundt wrote their dissertations in Germany, all of the following generations wrote their dissertations in the United States, predominantly at Columbia University, Harvard University, the University of Chicago and the University of Illinois. At the time, the United States, and especially the University of Chicago, had become the center of psychometrics with Louis L. Thurstone steering the wheel (see Figure 2).

James McKeen Cattell (Figure 3 and 4) was the first American to obtain a Ph.D. degree in psychology. His work on intelligence testing was strongly influenced by the versatile Francis Galton, who believed that traits were both heritable and measurable. Cattell finished his dissertation with Wundt in 1886, and in 1891, he accepted a professor position at Columbia University (where he was dismissed after speaking up against the World War I draft). This genealogy holds two of his graduate students: Edward L. Thorndike (Figure 3) and Robert S. Woodworth (Figure 4).

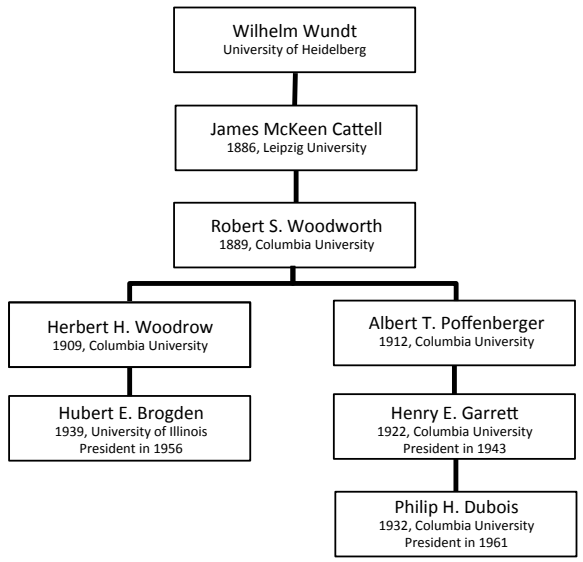


Figure 4. Genealogy of Wilhelm Wundt, part II. For each scholar in the genealogy, the year of obtaining a Ph.D. Degree and the university in question, are given. For presidents of the Psychometric Society, the year of their presidency is included.

The genealogy of Edward L. Thorndike

Edward L. Thorndike was a pioneer in learning theory, and among other things, investigated animal learning. Together with Thurstone and J. P. Guilford, Thorndike founded the Psychometric Society, and became its second president in 1936. He stayed at Columbia University for his entire career, and holds a central position in this genealogy, as the

advisor of four presidents: Jack W. Dunlap, Irving Lorge, Truman L. Kelley, and Edward E. Cureton. Note that Thorndike's son, Robert L. Thorndike, was already mentioned in Angell's genealogy (Figure 2).

Jack W. Dunlap was one of the driving forces behind the establishment of *Psychometrika* and he was president in 1940. He wrote his dissertation in 1931, which was a human factors study on the design of automobiles, and much of his career centered on industrial psychology. After working at a number of universities, he set up his own company, Dunlap & Associates, which was specialized in industrial consultancy. Irving Lorge, president in 1947, worked closely with E.L. Thorndike, especially in studies on language and communication. Together with Robert L. Thorndike (Figure 2), he developed the Lorge Thorndike Intelligence Test (1957). Edward E. Cureton, president in 1945, obtained his Ph.D. at the Thorndike laboratory in 1931, on the topic of measurement error. Throughout his career, he contributed to factor analysis and validity (Shrader, 1994).

Truman L. Kelley, president in 1938, was a psychometrician with a strong interest in both statistical methods and educational testing. He wrote books like *Statistical Method* (1923) and *Fundamentals of Statistics* (1947), but also cooperated with Lewis Terman (see Figure 6) on the Stanford Achievement Test Battery (1922). When Kelley became professor of psychology at Harvard University in 1931, he became the advisor of John C. Flanagan and Ernest Haggard. John C. Flanagan, president in 1951, became famous for his work on the first aptitude tests for American pilots in the Second World War, which he designed using a technique he called the 'critical incident technique' (Flanagan, 1954).

Ernest A. Haggard was the advisor of R. Darrell Bock at the University of Chicago, who wrote his dissertation in 1952. Bock, president in 1972, made significant contributions to multivariate analysis methods and IRT, and was the doctoral advisor of four presidents of the Psychometric Society: David Thissen, Robert Mislevy, Shizuhiko Nishisato and Ulf Böckenholt. The latter was the advisor of Alberto Maydeu Olivares, whose interest focus on structural equation modeling and item response theory, and who was president in 2013.

The fourth student of Thorndike in this genealogy was Herbert A. Toops. He worked primarily in the field of industrial psychology and designed tests for mechanical ability for trade schools. At Ohio State University, Toops advised two presidents: Harold A. Edgerton, in 1928, and Dorothy C. Adkins, in 1937. Edgerton, who was president in 1946, was also active in industrial psychology. Adkins was the first female president of the Psychometric Society in 1949. She finished her dissertation in 1937, worked as a test developer between 1940 and 1948, and returned to academia in 1948 at the University of North Carolina where she became professor of psychology (T. G. Thurstone, 1976). She was instrumental in bringing L. L. Thurstone there in 1952, after his retirement at the University of Chicago.

Toops was also the advisor of Robert J. Wherry, an industrial psychologist with strong interests in quantitative methodology. In 1951, he advised Ben J. Winer, who became president in 1967. Winer always worked for a psychology department but had strong interests in statistics, and his *Statistical Principles in Experimental Design* is one of the most cited works in psychological research (Haggbloom et al., 2002).

The genealogy of Robert S. Woodworth

From Cattell we find a second, shorter, lineage that starts with Robert S. Woodworth, who wrote his dissertation under Cattell in 1889. Woodworth and E.L. Thorndike investigated transfer of training; whether or not improvement of one function could also improve another function (Woodworth & Thorndike, 1901). From Woodworth, we find another T-junction: one pathway starts with Albert T. Poffenberger, the other with Herbert Woodrow.

Albert Poffenberger wrote his dissertation in 1912 on physiological psychology, and later in his career specialized in applied psychology. In 1922, he advised Henry E. Garrett, who was president of the Psychometric Society in 1943. Henry E. Garrett is most likely the most controversial president on the list. His work dealt mostly with intelligence, especially racial differences in IQ, and he did not shy away from putting his scientific views to political work: Garrett believed that racial differences in intelligence were innate and thus genetic (Garrett, 1961) and that mixing of races would have adverse consequences[4]; as a result, he helped organize a committee of researchers against racial mixing and wrote several papers in defense of racial segregation (Winston, 1998). He was the doctoral advisor of one president, Philip H. Dubois, who was a pioneer in machine test scoring, and worked in areas such as test construction and validity (Dubois, 1970).

Herbert Woodrow wrote his dissertation in 1909 at Columbia University[5]. His main interests were educational and experimental psychology. At the University of Illinois, he was the advisor of Hubert E. Brogden, who was president of the Psychometric Society in 1956. Most of Brogden's contributions were in the fields of military psychology and personnel selection.

The genealogy of Wilhelm Wundt, part III

Wundt's genealogy continues with four more graduate students (see Figure 5). Edward

[4] Important to note is that Henry Garrett was certainly not the only psychometrician who believed racial differences in intelligence were genetic. In the late 19th century and early 20th century, many scientists were supportive of Eugenics, the movement which aimed at improving the genetic quality of human beings (Norrgard, 2008).

[5] Woodrow's dissertation mentions Cattell, Thorndike, Lee and Woodworth as professors that have supported him throughout his time as a doctoral student. He is mentioned here as Woodworth's student, but his dissertation is not conclusive in that regard.

Titchener wrote his dissertation in 1892 on binocular vision under Wundt's supervision. After finishing his dissertation, he joined Cornell University where he developed a psychology laboratory. Unlike Cattell, Hall, and James, Titchener warmed up to Wundt's methods. Similarly to Wundt, Titchener believed that psychology could be investigated through experiments, especially experimental introspection (Greenwood, 2015). Titchener advised one president of the Psychometric Society: Joy P. Guilford, who was the third president of the Psychometric Society in 1937. One of Guilford's most famous studies involves the Structure of Intellect theory, in which intelligence could be traced back to a number of mental abilities (Guilford, 1956).

Hugo Münsterberg was a German psychologist, who first studied under Wundt, and then joined William James at Harvard, where he became director of the psychology laboratory. His work was mostly applied, and he was a forerunner of professional psychology in the United States. Münsterberg was the advisor of Robert M. Yerkes, intelligence tester and eugenicist. Yerkes was famous for his research on the psychology and physiology of primates. During the First World War, he led the intelligence testing program, which resulted in the Army Alpha and Army Beta tests, for literate and illiterate recruits respectively (Hilgard, 1956). At Harvard University, Yerkes advised Melvin E. Haggerty, who started out as an animal researcher, but moved onto education later in his career. At the University of Minnesota, Haggerty was the advisor of Philip J. Rulon, president of the Psychometric Society in 1948. After finishing his dissertation in 1931, he joined Truman Kelley at Harvard University, and took to statistics and educational measurement.

Wundt's fourth student in this genealogy is educational psychologist Charles H. Judd. He became director of the School of Education in 1909 at Yale University, and chairman of the Department of Psychology in 1920, following up James R. Angell (Buswell, 1947). From Judd, we found two lineages. The first starts with Guy T. Buswell, an educational psychologist who became famous for his work on the psychophysiology of reading, and was a pioneer in the field of eye-tracking. Buswell spent the first thirty years of his career at the University of Chicago, where Lee J. Cronbach was his graduate student. Lee Cronbach, president in 1953, is one of the most cited psychologists in the scientific literature (Haggbloom et al., 2002): his 1951 paper on the reliability coefficient alone has been cited over 35000 times. He is also well known for his work on construct validity together with Paul Meehl (Cronbach & Meehl, 1955) and for developing generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963).

The second student of Judd is the educational psychologist, Adam R. Gilliland, who was the advisor of Allen L. Edwards, who wrote his dissertation in 1940 and became president in 1963. Edwards was known for his skills for developing tests, the best-known being the Edwards Personal Preference Schedule, a personality inventory (Edwards, 1954).

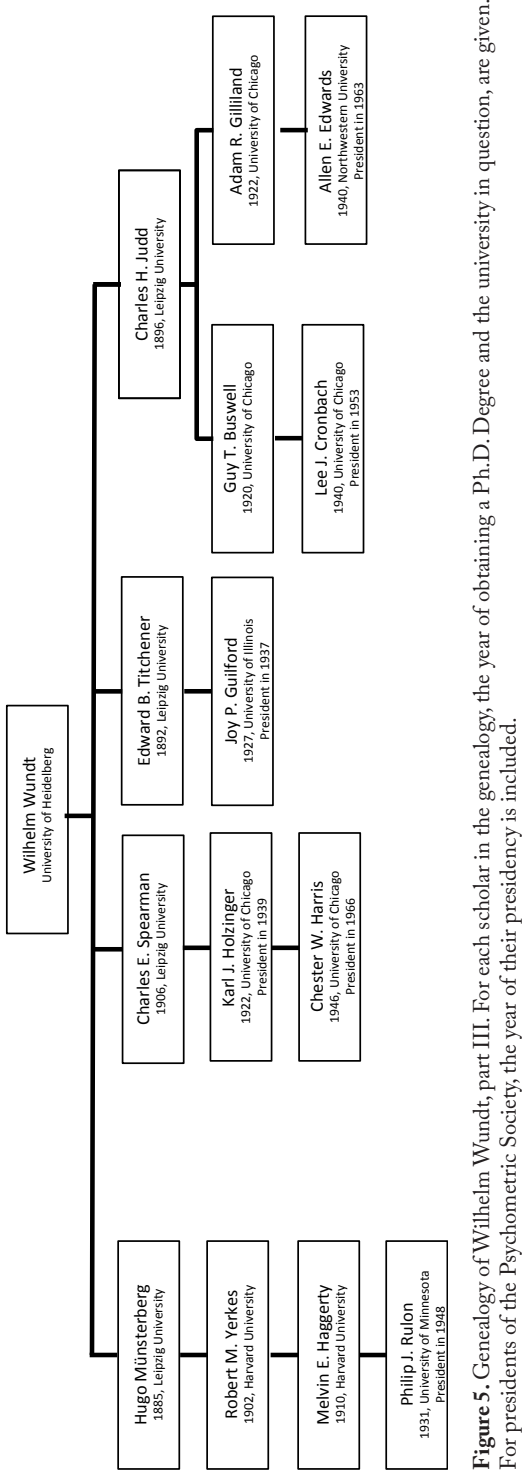


Figure 5. Genealogy of Wilhelm Wundt, part III. For each scholar in the genealogy, the year of obtaining a Ph.D. Degree and the university in question, are given. For presidents of the Psychometric Society, the year of their presidency is included.

With the exception of Wundt, we have so far only discussed psychologists and psychometricians who worked in the United States. One of the most influential psychometricians, however, was British: Charles Spearman (Figure 5), who wrote his dissertation in 1906 with Wundt. He became famous for his work on intelligence and the notion of *general intelligence*, or *g*, which he modeled with the first latent variable model: the common factor model (Spearman, 1904). His work on reliability and its relation to test length (the Spearman-Brown formula) marked the beginning of classical test theory (Traub, 1997). In this genealogy, Spearman has only one student in this genealogy, Karl Holzinger, the advisor of Chester W. Harris[6], who was president in 1966. Both were well known for their work on exploratory factor analysis.

2.3.3 The Genealogy of William James

William James has become known as the forerunner of American psychology. Including himself, his genealogy holds 16 scholars, of whom five became presidents of the Psychometric Society. James originally studied medicine, and in 1869 he earned his M.D. degree. James found that he was more interested in psychology and philosophy than medicine, but he never wrote a dissertation. It is well known that Wilhelm Wundt and William James diverged in their views on what kind of science psychology should be. Though James approved of Wundt's experimental method and certainly learned from it, he also found it boring and unnecessarily rigorous (Hilgard, 1987). So like James M. Cattell, Hugo Münsterberg and Charles H. Judd, James returned to the United States without the intention of continuing the European experimental tradition.

Instead, William James became famous for his pragmatism, a quintessentially American philosophy. According to pragmatism, theories should be primarily evaluated by their practical usefulness, and the pragmatic method can be used for "settling metaphysical disputes that otherwise might be interminable" (p.28, James, 1909). Rather than formulating psychological laws that describe what human responses take place when a certain stimulus is shown, like Wundt aimed to do, James focused on the *function* of psychological concepts; their purposes in life. Wilhelm Wundt, in turn, was disillusioned by the pragmatism of his American students, and thought that American psychology only focused on "inventions and physical comforts, rather than the deepening of man's understanding of himself and of nature" (Blumenthal, 1977, p.18). This conflict foreshadows the tension between research traditions that emphasize pragmatic usefulness versus those that emphasize scientific understanding, which is present in many fields,

[6] It is unclear whether Karl Holzinger was in fact Harris' official advisor. Holzinger is mentioned in the preface of Harris' dissertation as a main contributor, but so are Stephen M. Corey, Ralph W. Tyler, Paul B. Diederich and Osmond E. Palmer.

including psychometrics – an example is the discussion of interpretation of psychometric models as merely being a useful description of the data versus scientific explanatory models of test behavior (such as De Boeck & Wilson, 2004).

William James stayed at Harvard University, until he retired in 1907. Two of his students can be found in this genealogy: Granville Stanley Hall and Morris R. Cohen. In 1878, it was G. Stanley Hall who was the first to obtain a Ph.D. degree in psychology awarded by an American university, under James' supervision. Hall also spent a short time at Wundt's laboratory, and like many of Wundt's other students, Hall did not carry over many traditional German ideas on how to do psychological research. Hall established the first American experimental psychology laboratory at John Hopkins University, and was initiator of the American Psychological Association in 1892. G. Stanley Hall was a pioneer in the field of educational psychology and believed that psychological research should serve a practical purpose and be applied to real-world problems.

From G. Stanley Hall, we found two separate lineages. The first one starts with Joseph Jastrow. In 1888, only two years after finishing his dissertation on spatial perception in vision and touch at John Hopkins University, he set up a psychological laboratory at the University of Wisconsin (Hull, 1944). During his career, Jastrow actively sought out the attention of the public, and wrote many popular publications besides his scientific work. One of Jastrow's students at Wisconsin University was Clark L. Hull, known for the reinforcement theory of learning. Hull was a strong quantitative thinker and in all stages of his research, he strove for solid theories and a certain mathematical rigor.

Upon James Angell's request, Hull took a position at Yale University in 1929. There, he headed a group of researchers who developed a research program devoted to the principles of learning, using rat experiments (Greenwood, 2015). Two of his graduate students are in this genealogy, Neal E. Miller and Carl I. Hovland (Figure 6). Initially, Miller made an effort to experiment on Freudian ideas, like fear or frustration, and argued that behaviorist notions like conditional learning also applied to those. In the sixties, he also took up brain studies, and became one the biggest names in the new interdisciplinary field of 'biofeedback'. Neal E. Miller was the advisor of Gordon Bower, now emeritus cognitive psychologist. Bower left Yale University after finishing his dissertation in 1959, and went to Stanford University. For his dissertation, Bower used animal experiments to investigate reinforcement mechanisms, but when he left Yale, he concentrated more on mathematical models for memory. Bower was the advisor of Phipps Arabie, who was president of the Psychometric Society in 1990. Arabie was active in the fields of multidimensional scaling, clustering and combinatorial data analysis (Hubert et al., 2001). He became founding editor and guiding spirit of the *Journal of Classification*, the flagship journal of the Classification Society of North America (Hubert, 2012).

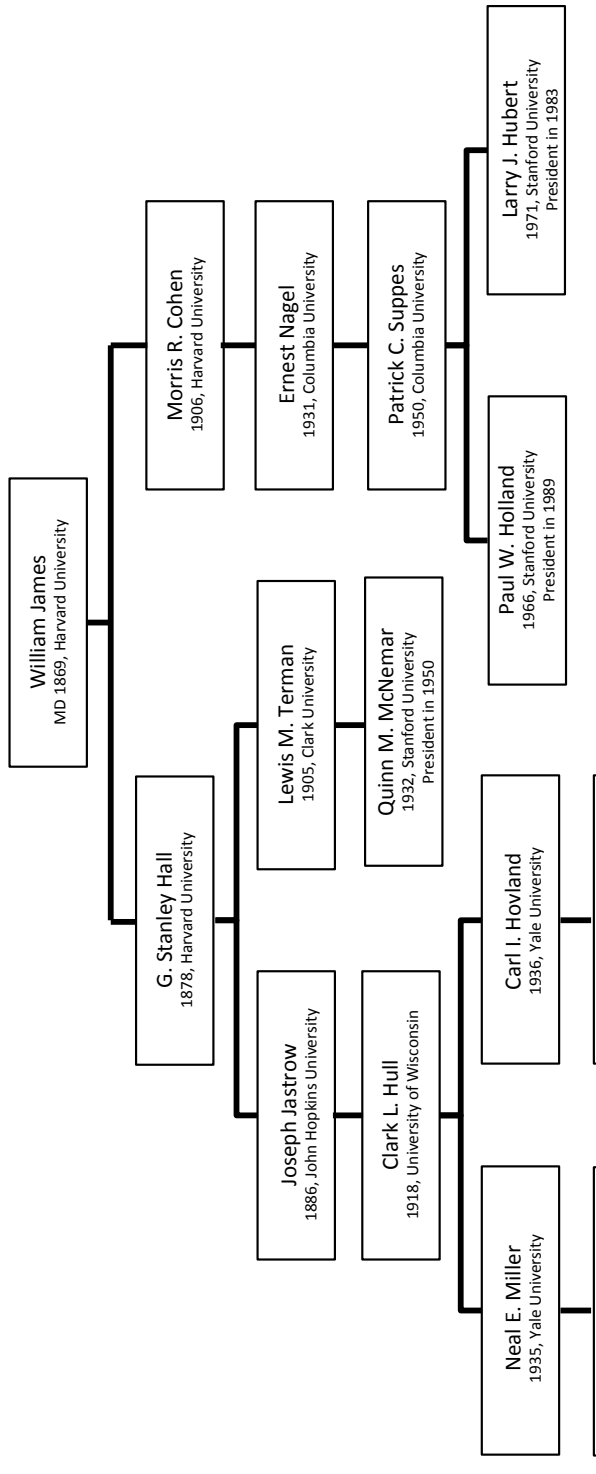


Figure 6. The genealogy of William James. For each scholar in the genealogy, the year of obtaining a Ph.D. Degree and the university in question, are given. For presidents of the Psychometric Society, the year of their presidency is included.

Another student of Hull in this genealogy was Carl I. Hovland. Hovland, who wrote his dissertation in 1936, always showed a wide interest in many topics, and applied the experimental method to all sorts of problems. One of his best-known research projects took place during WWII, when he investigated the effectiveness of information programs on the motivation of American soldiers (Sears, 1961). Hovland was the advisor of Roger N. Shepard, president of the Psychometric Society in 1973. Shepard was a cognitive psychologist who belongs to the list of the 100 most eminent psychologists of the 20th century (Haggbloom et al., 2002), especially because of his work on spatial relations and mental rotation. That work also led him to consider nonmetric approaches to multidimensional scaling, which inspired later presidents like Joe Kruskal, Forrest Young, Jim Ramsay, Yoshio Takane and Jan de Leeuw in their groundbreaking nonmetric approaches to multidimensional data analysis.

So far we have discussed the first lineage that starts with G. Stanley Hall at John Hopkins University. In 1888, G. Stanley Hall left John Hopkins University and became the first president of Clark University, a new university mostly dedicated to graduate education in psychology. There, he became the doctoral advisor of Lewis M. Terman, who wrote a dissertation on testing methods in 1905, even though Hall discouraged him from doing so (Hilgard, 1987). From 1922 until 1942 he was a professor of psychology at Stanford University, where he worked on what would become his most famous work: the Stanford Binet and the IQ test. One of his Ph.D. students was Quinn McNemar, who was president in 1950. McNemar developed a revision of the Stanford-Binet Scale in 1942, and wrote his famous book *Psychological Statistics* (1949).

A second student of William James was Morris R. Cohen, who was prominent in a wide variety of domains, e.g. law, education and philosophy. His student, Ernest Nagel, became a famous philosopher of science. Nagel's 1931 thesis was on the foundations of measurement, which continued to be an important theme in Nagel's career. Other recurring topics were causality, scientific method and explanation, the foundations of probability, and induction (Suppes, 1994). Ernest Nagel stayed at Columbia University for most of his career, where he was the advisor of Patrick Suppes, who finished his dissertation in 1950. Like his advisor, Suppes was a versatile scholar: he was well known as philosopher of science, but was also involved in experimental psychological research and founded a company that provided educational software (Moulines, 2016). After finishing his dissertation, he left for Stanford University where he advised many students, including two presidents of the Psychometric Society: Paul W. Holland and Larry Hubert. Paul W. Holland was active in many domains of the social sciences, such as educational testing at ETS, social networks and causal inference. Larry Hubert is especially known for his work on graphs, trees, and clustering. He was editor of *Psychometrika* from 1988 to 1992. In close collaboration with later presidents Phipps Arabie and Jacqueline Meulman he

developed the new field of combinatorial data analysis (Hubert, Arabie & Meulman, 2001).

2.3.4 The Genealogy of Albert E. Michotte

Figure 7 displays the genealogy of Albert E. Michotte, a Belgian psychologist from Leuven University, who is the ancestor of four Dutch and two Belgian presidents. Though Albert E. Michotte also spent time with Wundt in Leipzig, Wundt was not his doctoral advisor. From Michotte, we found two pathways, one of which starts with the Dutch psychologist, Franciscus J. M. A. Roels who went to Utrecht University after finishing up his Ph.D. degree in Leuven. He was well known for developing psychotechnic apparatus for the selection of employees. At Utrecht University, he was the advisor of F. J. Theo Rutten. Rutten became professor of psychology at Nijmegen University, and set up the psychology laboratory. His student, Alfons Chorus, was an assistant in that same lab. Both Rutten en Chorus believed that psychology could affect the world in a positive way and contribute to a more pleasant society. In 1947, Chorus became the founding professor of psychology at Leiden University.

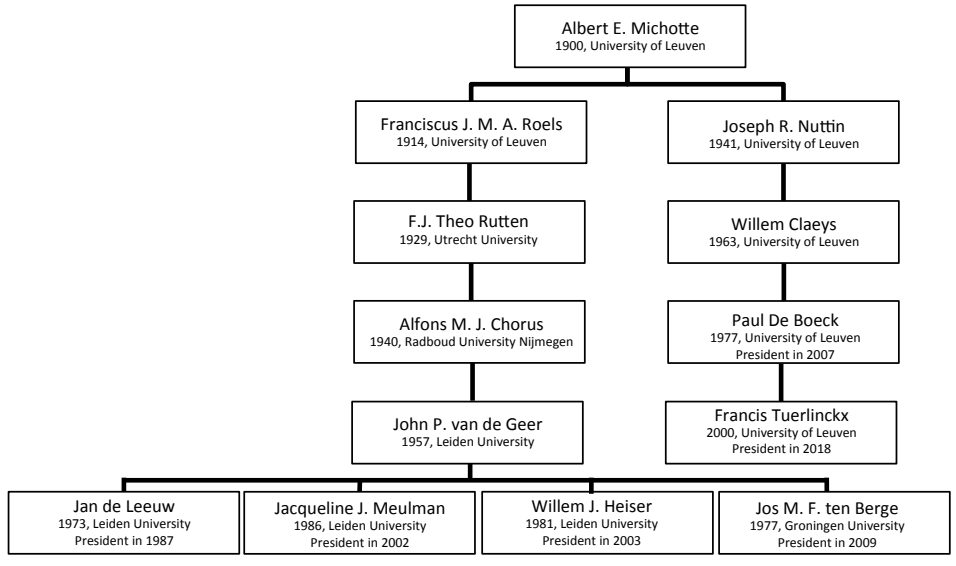


Figure 7. The genealogy of Albert E. Michotte. For each scholar in the genealogy, the year of obtaining a Ph.D. Degree and the university in question, are given. For presidents of the Psychometric Society, the year of their presidency is included.

In 1957, John van de Geer received his Ph.D. degree under Chorus' supervision. He was a cognitive psychologist who published one of the first sophisticated applications of nonmetric MDS (Levelt, Van de Geer, & Plomp, 1966) and wrote an influential textbook

about multivariate analysis (Van de Geer, 1971; cf. Heiser, 2008). He founded the highly regarded department of Data Theory in Leiden. Four of his students became president of the Psychometric Society: Jan de Leeuw, Jacqueline Meulman, Willem Heiser, and Jos ten Berge (see Figure 7). The first three integrated and extended the nonmetric techniques pioneered by Shepard, Kruskal, and Guttman, publishing with their co-workers under the pseudonym Albert Gifi (cf. Van der Heijden & Sijtsma, 1996). Ten Berge became known for his work on reliability and factor rotation, but especially for his contributions to three-way data analysis, in the footsteps of presidents Tucker, Carroll, and Kruskal.

The second branch that departs from Michotte starts with Joseph Nuttin, who, besides receiving his degree in 1941 from Michotte, was also an ordained priest. His main interests were learning, motivation and personality. He stayed in Leuven for his entire career, where he supervised Willem Claeys, a personality psychologist. Claeys was the advisor of Paul De Boeck, who is now professor of psychology at Ohio State University and was president of the Psychometric Society in 2007.

De Boeck's work is characterized by a strong interest in individual differences and quantitative approaches, among which is explanatory item response models. Explanatory IRT exemplifies the idea that psychometric models are not merely technical vehicles for data analysis, but can also be used to build explanatory models of test behavior (De Boeck & Wilson, 2004). His student, Francis Tuerlincx, has become president of the Psychometric Society in 2018. Tuerlinckx' work is strongly affiliated with mathematical psychology, and combines important models taken from mathematical psychology (e.g. the diffusion model) with psychometric models (e.g. the IRT model; see Tuerlinckx & De Boeck, 2005).

2.3.5 The Genealogy of Carl Friedrich Gauss

The genealogies mentioned above have clear roots in psychology, and although the branches are separate when defined in terms of doctoral advisors of advisees, there are several occasions when people of separate genealogies met and even cooperated. However, not all presidents from the Psychometric Society have ancestors in psychology: an important subset of presidents, mostly of strong mathematical inclination, descends from the German mathematician Carl Friedrich Gauss. It is interesting to observe that two of the important ingredients of psychometrics – psychological research and mathematical analysis – are also visible in the form of distinct genealogies that can be traced back to psychology and mathematics. The genealogy of Gauss holds 40 scholars, including 9 presidents.

Carl Friedrich Gauss, being one of the most important mathematicians in history, contributed to a wide variety of fields, and Jones and Thissen (2007) consider his and Friedrich Wilhelm Bessel's work (Bessel is a descendent of Gauss, see Figure 8) work in astronomy as one of the cornerstones of psychometrics. He showed that the accumulation

of small random deviations leads to a so-called Gaussian –or normal- distribution. The resulting 'theory of errors', which involves the construction of a mathematical model that captures the behavior of measurement error, can also be considered the motivation for basic axioms in test theory (e.g., the notion that scores can be decomposed in a true score an error component as well as the idea that errors are uncorrelated and give rise to symmetric bell-shaped distributions; Lord & Novick, 1968). In addition, Bessel's work was the first to give an analysis of differences between observers who recorded measurements of the exact same stimuli (i.e., positions of stars). Bessel developed the 'personal equation', which corrected observations for systematic effects induced by different observers, and in doing so gave one of the first formal approaches to the notion of *measurement bias*, which has also become a fundamentally important concept in test theory (Lord, 1980; Mellenbergh, 1989; Meredith, 1993).

As Figure 8 shows, Gauss' genealogy incorporates many mathematicians whose relevance to psychometrics is tangential. Describing each of them goes beyond the scope of this paper, so this section will only focus on the presidents who descended from Gauss.

Frederick Mosteller, president in 1957, and a student of Samuel S. Wilks, was a statistician with a profound interest in mathematical psychology, which resulted in a book *Stochastic Models for Learning* (1955), co-written with physicist Robert R. Bush. Over the course of his career, he collaborated with many prominent statisticians, among who are Joseph B. Kruskal (see Figure 5) and John Tukey (Fienberg, Hoaglin & Tanur, 2013). Kruskal, a student of Roger C. Lyndon and president in 1974, was a mathematician who, in the psychometric community, was most active in the domains of MDS and three-way data analysis. Besides his theoretical work, he also aimed to apply MDS to a variety of topics, like lexicostatistics or glottochronology (Carroll & Arabie, 2013). In 1958, he moved to Bell Labs in New Jersey, where he worked in research until his retirement.

William F. Stout, retired professor of statistics at the University of Illinois (where he stayed throughout his entire career), was president in 2001. Stout was a professor of mathematics for a number of years before he made the switch to statistics and psychometrics. In psychometrics, he made important contributions to Differential Item Functioning (DIF), unidimensionality and diagnostic classification. He was the doctoral advisor of three presidents of the Psychometric Society: Brian W. Junker, president in 2008, Hua-Hua Chang, president in 2012, and Daniel Bolt, president-elect at the time of writing.

Melvin R. Novick, student of William J. Hall, was already mentioned as one of the two famous authors, the other being Fred Lord, of *Statistical Theories of Mental Testing* (1968). He remained active in the world of testing, and became a frequent advisor at ETS. He was a pioneer of the use of Bayesian statistics in psychometrics (Lindley, 1987) and co-designed a statistical computing system, the CADA monitor (Novick, Hamer, & Chen, 1979).

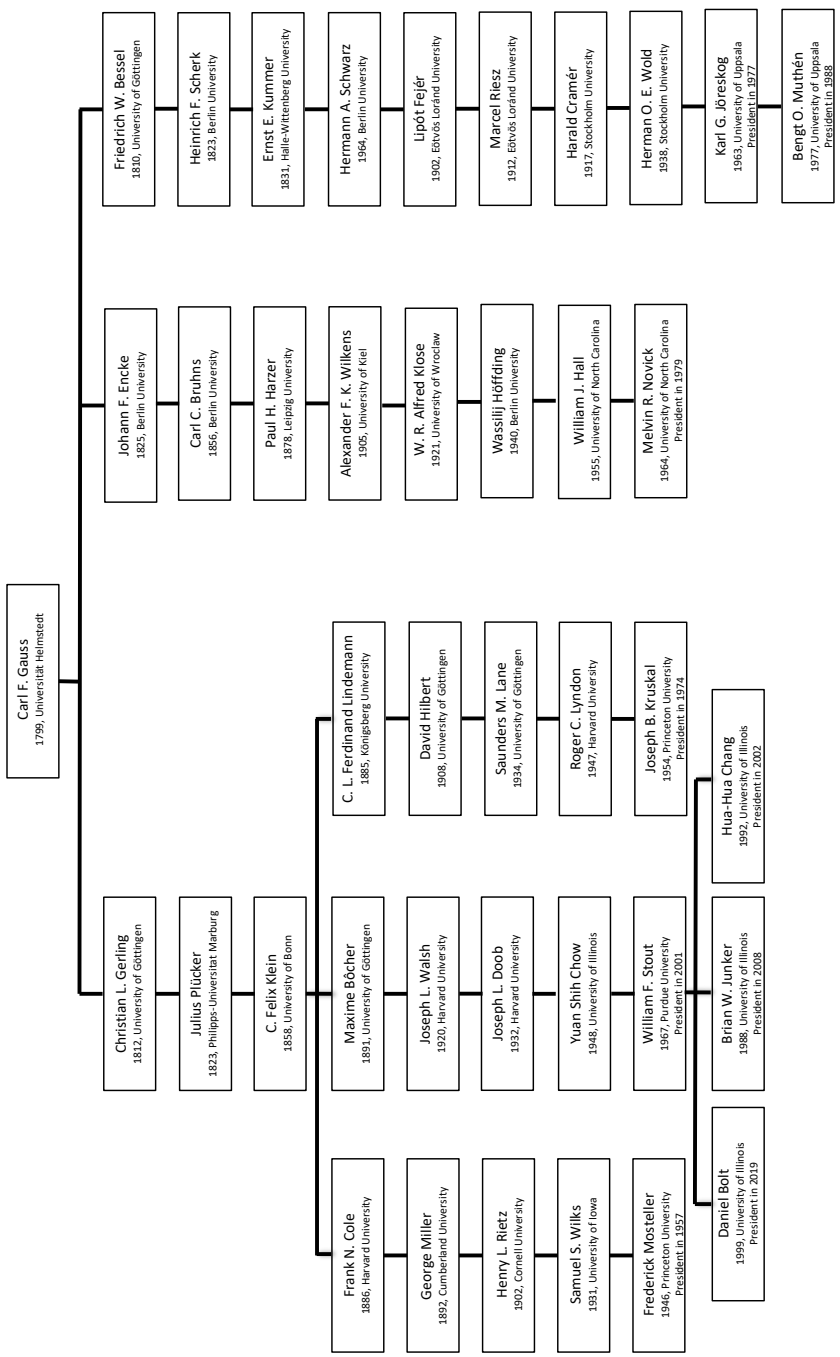


Figure 8. The genealogy of Carl Friedrich Gauss. For each scholar in the genealogy, the year of obtaining a Ph.D. Degree and the university in question, are given. For presidents of the Psychometric Society, the year of their presidency is included.

The last two presidents that connect to the Gauss genealogy are Karl G. Jöreskog and Bengt O. Muthén. Jöreskog was president in 1977, and became one of the most influential psychometricians of the second half of the 20th century. After finishing his dissertation in Uppsala in 1963, he left for ETS, working together with Frederick Lord and Melvin Novick (Wainer, 2011). He returned to Sweden in 1971 to become one of the forerunners of Structural Equation Modeling (SEM). Together with Dag Sörbom, he developed the statistical program LISREL – the first widely available psychometric software program to become a success throughout the sciences. Jöreskog supervised many Ph.D. students, among was Bengt Muthén, who became president in 1988. Muthén is now professor emeritus at UCLA. He is one of the developers of the computer program *Mplus*, which is specialized in latent variable modeling and implements many important extensions of the structural equation model.

2.3.6 Lineages of Other Presidents

Including the president-elect of 2018-2019, there are 84 presidents of which 20 presidents cannot be connected to any of the genealogies above. In this section, we discuss their lineages, as represented in Table 1. The table shows the lineages for each of these presidents separately, including the year they wrote their dissertation, the university where they were a graduate student and their year of presidency.

Some of the presidents (e.g. Ivo W. Molenaar, R. Duncan Luce, Michael W. Browne) in this table have a lineage completely separate from other presidents. There are also a number of advisor-advisee couples in this table, for which one lineage holds two presidents. For example, Cees Glas, president in 2017, was the student of Willem van der Linden (president in 1999), Lyle V. Jones, president in 1962, was the advisee of Lloyd G. Humphreys (president in 1959), and Klaas Sijtsma, president in 2010, was the student of Ivo W. Molenaar, president in 1997.

Surprisingly, the lineage of Jones and Humphreys, both American psychometricians, and the lineage of Gerhard Fischer, psychometrician from Vienna University, share a common ancestor: Benno Erdmann. Benno Erdmann was a German Neo-Kantian philosopher, who held positions at different of German universities. At Halle University, he advised Raymond Dodge, an American philosophy student who was not admitted to Harvard or Columbia and crossed the Atlantic instead (Miles, 1956). Dodge returned to the United States, and specialized in experimental psychology. At Yale University, he advised Ernest Hilgard, the advisor of Humphreys. The second branch from Erdmann goes via Erich Becher, Hubert Rohracher and Erich Mittenecker to Gerhard Fischer, a specialist in IRT, who became president in 1994.

Table 1. Year and University of Graduation, Year of Presidency and Lineages of 'Unconnected' Presidents of the Psychometric Society.

Name of president	Year of graduation	University where dissertation was written	Year of presidency	Lineage
Cees Glas	1989	University of Twente	2017	Wim J. van der Linden - Gideon J. Mellenbergh – Adriaan D. de Groot – Géza Révész – Georg E. Müller – Hermann Lotze
Sophia Rabe-Hesketh	1992	King's College, University of London	2014	James F. Boyce
Anders Skrondal	1996	University of Oslo	2013	Petter Laake
Mark R. Wilson	1984	University of Chicago	2011	Benjamin D. Wright – Bruno Bettelheim – Robert Reininger
Klaas Sijtsma	1988	University of Groningen	2010	Ivo W. Molenaar – Jan Hemelrijk – David van Dantzig – Bartel L. van der Waerden – Hendrick de Vries – Diederik J. Korteweg
Wim J. van der Linden	1980	University of Amsterdam	1999	Gideon J. Mellenbergh– Adriaan D. de Groot – Géza Révész – Georg E. Müller – Hermann Lotze
Susan E. Embretson	1973	University of Minnesota	1998	Rene Dawis – Donald G Paterson
Ivo W. Molenaar*	1970	University of Amsterdam	1997	Jan Hemelrijk – David van Dantzig – Bartel L. van der Waerden – Hendrick de Vries – Diederik J. Korteweg
Fumiko Samejima	1965	Keio University	1996	Taro Indow
Gerhard H. Fischer	1963	Vienna University	1994	Erich Mittenecker – Hubert Rohracher – Erich Becher – Benno Erdmann
Michael W. Browne*	1968	University of South Africa	1991	Hendrik S. Steyn – Alexander C. Aitken – Edmund T. Whittaker – Andrew R. Forsyth – Andrew Cayley
Roderick P. McDonald	1963	University of New England	1985	John Keats
Peter M. Bentler**	1964	Stanford University	1982	Douglas Jackson – John M. Hadley – John R. Knott – Lee E. Travis – Carl E. Seashore – George T. Ladd
R. Duncan Luce*	1950	Massachusetts Institute of Technology	1976	Irvin S. Cohen – Oscar A. Zariski – Guido Castelnuevo – Giuseppe Veronese – Antonio L. G. G. Cremona

Name of president	Year of graduation	University where dissertation was written	Year of presidency	Lineage
Louis Guttman	1942	University of Minnesota	1970	F. Stuart Chapin – Franklin H. Giddings
Henry F. Kaiser	1956	University of California	1969	Harold D. Carter – Donald G. Paterson
Harry H. Harman			1968	Never wrote a dissertation
Lyle V. Jones	1950	Stanford University	1962	Lloyd G. Humphreys - Ernest Hilgard – Raymond Dodge – Benno Erdmann
John B. Carroll	1941	University of Minnesota	1960	B. F. Skinner – William J. Crozier – Jacques Loeb
Lloyd G. Humphreys	1938	Stanford University	1959	Ernest Hilgard – Raymond Dodge – Benno Erdmann

* The entire lineages of Ivo W. Molenaar, Michael W. Browne and R. Duncan Luce can be found on the website of the Mathematics Genealogy Project; in this table we only mention five of their ancestors.

** Since the publication of this article, it has come to light that Peter Bentler's genealogy can perhaps be traced back to Wilhelm Wundt. According to this information, Lee E. Travis is not a descendent from Carl E. Seashore but from J. J. B. Morgan instead, who was a student of Woodworth's (who is part of James McKeen Cattell's genealogy). Note that this information is not yet verified.

The individual lineages of Susan Embretson, president in 1998, and Henry F. Kaiser, president in 1969, also share a common ancestor: Donald G. Paterson. He was a pioneer in applied psychology, especially in the field of vocational guidance. Though he studied with Rudolph Pintner, a former student of Wilhelm Wundt, he never officially wrote his dissertation.

Table 1 also shows that for a number of presidents (Sophia Rabe-Hesketh, Anders Skrondal, Fumiko Samejima and Roderick McDonald), we have only included their own advisor. The remainder of their lineages are simply unknown to us. It is unlikely though that their lineages will connect to any of the other existing lineages, as they either come from different fields (Rabe-Hesketh has a background in physics, Skrondal in biostatistics) and/or have received their training in different countries that are not strongly represented in this genealogy (resp.: United Kingdom, Norway, Japan and Australia). Therefore, we have refrained from further research into their genealogies.

Louis Guttman, president in 1970, is an important outsider in this genealogy. He was born in New York, received his Ph.D. training in quantitative sociology at the University of Minnesota with his advisor, F. Stuart Chapin, and graduated in 1942. He then immigrated to the new state of Israel in 1947, where he became the founding director of the Israel Institute of Applied Social Research. Guttman gained an outstanding reputation as a psychometrician, due to his contributions in the American Soldier project during the Second World War, his work on exploratory factor analysis and MDS, and most famously perhaps, to the deterministic scaling model that carries his name (Guttman, 1944).

Mark R. Wilson was the graduate student of Benjamin D. Wright, who was a famous psychometrician in the Rasch model tradition. Wilson was president in 2011 and collaborated with Paul De Boeck on explanatory item response theory (De Boeck & Wilson, 2004). Wilson's focus lies with a wide variety of measurement and assessment issues, and is now professor at the Graduate School of Education of UC Berkeley.

Peter M. Bentler, president in 1982, was the doctoral student of Canadian psychologist Douglas N. Jackson. Bentler, though trained as a clinical psychologist, is especially known for his work on SEM and the development of EQS, a software package for SEM modeling (Bentler, 1995). He has also contributed to the fields of personality and drug abuse.

John B. Carroll, president in 1960, wrote his dissertation under B. F. Skinner, the famous behaviorist, at the University of Minnesota, but during his time as a graduate student he spent time at Thurstone's lab at the University of Chicago. Thurstone had a strong influence on him, perhaps even more than Skinner, and Carroll ended up writing his dissertation on factor analysis of verbal abilities. Throughout his career, he made important contributions to the fields of intelligence research and testing.

Harry Harman is the only president without a doctorate degree and is therefore without a lineage. However, he made many contributions to the field. In 1936 he obtained a Master's degree from the University of Chicago, which at the time was the center of psychometric research (see Figure 2). He closely collaborated with Karl Holzinger in many articles and co-wrote a handbook on factor analysis (Holzinger & Harman, 1941).

2.4 Conclusion & Discussion

We have developed a systematic overview of the lines of descent of Psychometric Society presidents through the construction of an academic genealogy. The graphs provide a broad overview of the people involved in these lines of descent and how they are interconnected; in addition, the genealogy is the first of its kind that provides verifiable evidential backup for each of the connections in the system. The genealogies we have reported have first and foremost a strongly descriptive purpose, as they describe the sequence of people in several individual lines of descent and identify common ancestors. Since the historical research necessary to document the relevant relations becomes more complicated as time goes on, the current paper also has an important function of preserving and disseminating this historical knowledge.

Besides these descriptive and preservationist purposes, genealogies can also serve as stepping-stones for further hypothesizing about the history of psychometrics. In the following, we present hypotheses about the history of psychometrics that are derived from one or more of the graphs.

2.4.1 Missing Offspring

One of the questions that this genealogy raises is why certain scholars have produced a number of presidents, whereas other scholars, some of whom were equally if not more prominent in psychometrics, have very little or no offspring at all in the genealogy. An interesting example of an important name in psychometrics that has not had that same productivity within the Psychometric Society is Lee J. Cronbach. Cronbach is probably one of the most well-known psychometricians and is certainly the best cited psychometrician in psychology at large (Ho & Hartley, 2016). It is therefore quite surprising that Cronbach is a terminal node in the genealogy presented here.

Another influential psychometrician who is not prominently present in the genealogy is Charles Spearman. Spearman is probably one of the most important psychometricians that ever lived, and his work has marked the birth of psychometrics as a discipline. However, unlike Thurstone's, Spearman's lineage in this genealogy is rather short. So why is it that Thurstone's 'formal' influence reaches so much further than Spearman's? One of the reasons might be that it was Thurstone who formalized the discipline by setting up the Psychometric Society and *Psychometrika* in the United States. Thurstone set up his own laboratory, first in Chicago, and later on at the University of North Carolina, and collected a large group of people that shared his ideas and developed them further (see Figure 1), but Spearman never officially connected to this group. And though Thurstone and his followers were greatly inspired by Spearman and shared an interest in measuring mental abilities, they disagreed on some fundamental points, one being the number of mental abilities. So not only was Spearman not part of the Chicago-group, his and Thurstone's work were simply incompatible (Thompson, 1947).

A number of influential psychometricians are entirely missing from this genealogy, such as Gustav Fechner, Francis Galton, and Raymond B. Cattell. The genealogy does not give an integrative explanation why some psychometricians are missing from the genealogy, and others are only marginally present. There are simply too many factors that can play a role: geography, interpersonal conflict, involvement with other groups, or simply a lack of graduate students who were active within the Psychometric Society. What these examples do show is that the number of descendants does not have a one-to-one relation with the extent to which a psychometrician was successful or influential: but rather with the extent to which a psychometrician was influential in the *formalization* and the *institutionalization* of psychometrics as an independent scientific discipline, and specifically, in the institutionalization of the Psychometric Society. So interpreted, our genealogy shows that these two kinds of importance do not necessarily coincide. This may be an important finding in the light of the way psychometricians portray the history of their field; psychometrics as an institutionalized discipline is more the field of Thurstone, Gulliksen, and Bock (those that trained students who also became presidents) than that of Spearman and Cronbach.

2.4.2 Disciplinary Boundaries

The results section shows that the majority of psychometric presidents can be divided into a number of genealogies rooted in psychology, and one genealogy that is rooted in mathematics. These findings suggest that psychometrics is an inherently multidisciplinary field, which embodies distinct traditions coming out of mathematics (specifically, statistical analysis) and psychology (specifically, psychological measurement and educational testing). We suggest that this inherent multidisciplinarity is responsible for the considerable tension between substantive and technical orientations that has played an important role in the history of Psychometric Society.

This tension has been particularly visible in several splits that have occurred in the history of psychometrics, and that have led to the formation of at least two distinct professional organizations that deal with topics that were originally seen as belonging to the realm of psychometrics (Green, 1986). First, during the history of psychometrics, a sharp line was drawn between psychometrics and mathematical psychology. In the 1960s, several mathematical psychologists such as Duncan Luce and Patrick Suppes founded the Society for Mathematical Psychology and its flagship journal, *The Journal of Mathematical Psychology*, thereby formally separating mathematical psychology from psychometrics. Ever since, the two have existed in distinct realms.

Another important and very similar schism took place with the formation of the Society for Experimental Psychology (SMEP) in 1960 and its journals *Multivariate Behavioral Research* and the less-known *Multivariate Experimental Clinical Research* in 1966. Although this event appears less clearly related to psychometrics as a discipline, in Cattell's (1990, p. 49) recollection of the history of SMEP he does state that "members should be chosen on the basis of experimental, substantive work, not of virtuosity in statistical ideas", which appears to echo the concern that formalized psychological work received scant attention in psychometrics. Green (1988) in fact relates this event specifically to the editorial policy of *Psychometrika*, which was to only publish substantive work if it served the purpose of illustrating a larger psychometric theme, and not for its own sake.

The lack of integration of psychology and psychometrics thus appears to be at least partly a problem of the Psychometric Society's own making. Attempts to remedy this situation, for instance by including an Applications section in *Psychometrika*, have been only partly successful. The present study suggests that the tension between psychology and mathematics, may in fact run deep: the Psychometric Society is in essence a combination of disciplines of different origins, which bring with them different investigative styles and research traditions. As a function of these repeated splits, however, psychometrics has been increasingly dominated by statistically oriented rather than psychological approaches, so much so that many psychometricians today see psychometrics as a branch of statistics rather than a branch that is "devoted to the development of psychology as a quantitative

rational science" (the original motto of *Psychometrika*, appearing on its masthead until 1984). This trend is in contrast to Thurstone's (1937, p. 232) first presidential address to the society, in which he said:

> Let us remember that a psychological theory is not good simply because it is cleverly mathematical, that an experiment is not good just because it involves ingenious apparatus and that statistics are merely the means for checking theory with experiment. In the long run we shall be judged in terms of the significance, the fruitfulness and the self-consistency of the psychological principles that we discover.

2.4.3 Increasing Diversity

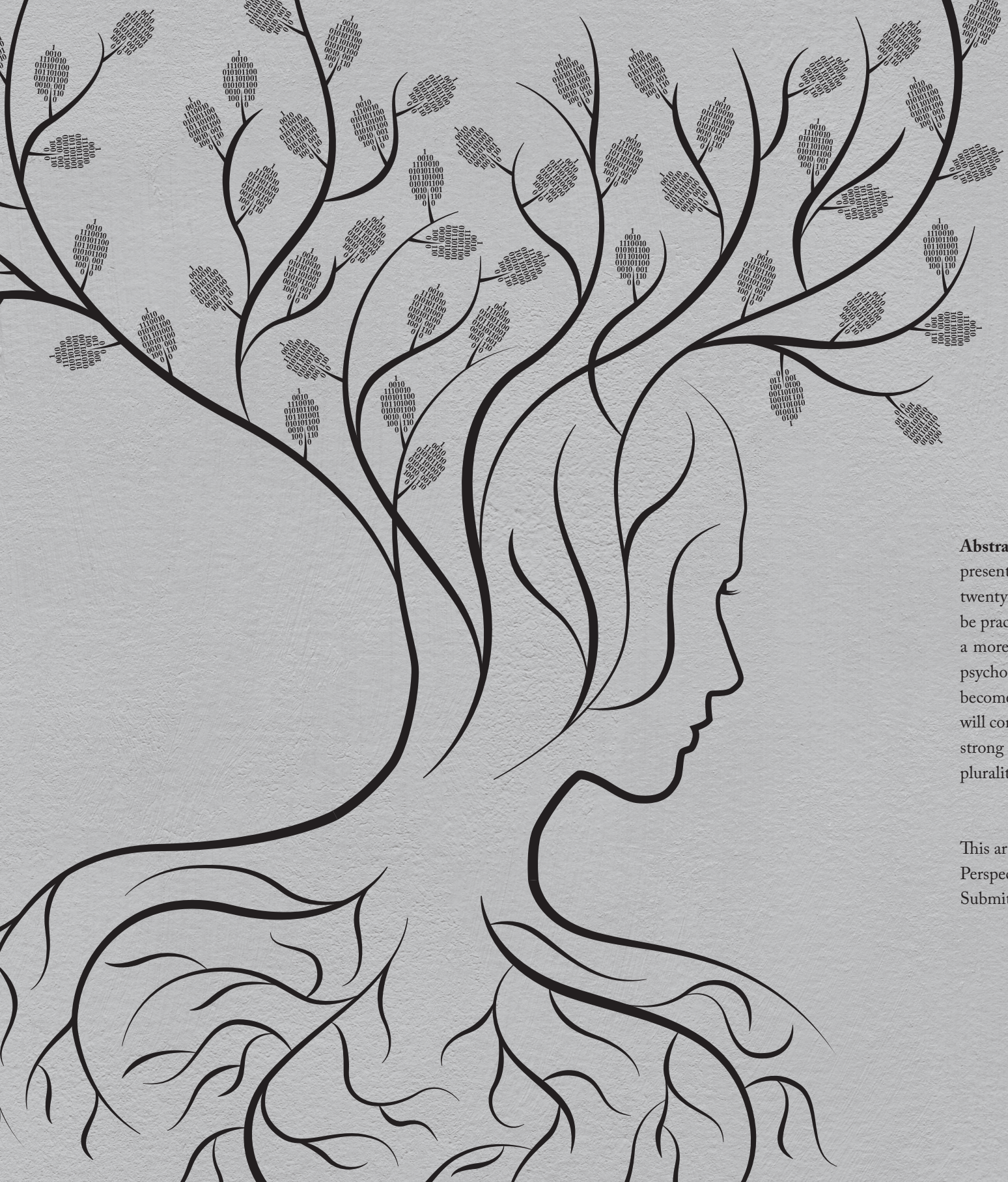
The Psychometric Society was originally an American institution, and the majority of the Presidents have been North American and received training at one of the major universities there. The first president who did his graduate education outside of North America was Karl Jöreskog. The meeting he organized, in 1978 in Uppsala, Sweden, was also the first meeting of the Psychometric Society outside the United States or Canada. After Jöreskog's presidency, many other non-Americans followed, and meetings in the Netherlands, Belgium, France, and many other countries have been organized since. In 2001 the first Asian meeting was organized in Osaka, and in 2019 the first South-American IMPS will be held in Chile. A second increase in diversity that is evident in the history of the Psychometric Society relates to gender: we count four female presidents in the past 25 years (i.e. in the period 1993-2018), which does not seem like much. However, it represents a spectacular increase relative to the long period that went before (1935-1993), which counted only one female president (Dorothy Adkins, president in 1949). Clearly, there is an increase in diversity there as well. Interestingly, a similar process may operate at the level of substantive backgrounds of psychometricians, as judged by the increasing number of presidents that have received a different type of training from that of the early presidents: there is a sizable influx of presidents with a mathematics background, mostly in the later stages of the Psychometric Society's history (see the Gauss genealogy, Figure 8). This indicates that the development of psychometrics is characterized by an increasing diversity, both in terms of socio-cultural and gender composition and in terms of substantive background.

Both the increasing diversity, and the formation of the breakaway groups, also suggest that the Psychometric Society is now possibly less representative of the psychometric community than in its early days, when the psychometrics community was significantly smaller and centered around a small number of universities. Psychometrics has become a widely practiced discipline, and the Psychometric Society, though still of high importance for the psychometric community, no longer represents it entirely.

2.4.4 Limitations

With the selection of the presidents of the Psychometric Society as the initial set of nodes in the genealogy come a number of limitations. By making a selection, we have excluded important or influential psychometricians who were never president. Moreover, this genealogy only shows how ideas may be spread through advising students, and loses sight of other ways psychometricians have been influential, such as being a productive author or a meeting organizer. We want to stress that we by no means intend to provide a complete account of the history of psychometrics, and that this genealogy only sheds light on parts of it. Nevertheless, this study offers an important contribution to the literature on the history of psychometrics. The presidents of the Psychometric Society have each fulfilled important roles in the development of psychometrics and their lines of descent are thus of importance to the history of psychometrics. Further research is needed to do justice to the complexity of the history of psychometrics as a whole.

In line with the argument given above, groups that engage in psychometric research but that are not strongly affiliated with the Society might find little or no representation in this genealogy, despite the fact that they may have played important roles in other ways. Future research could try to uncover these groups by a co-author network analysis of publications in *Psychometrika* and in related journals like the *Journal of Educational and Behavioral Statistics*, *The Psychological Bulletin*, *Psychological Methods*, the *British Journal of Mathematical and Statistical Psychology*, *Multivariate Behavioral Research*, *Structural Equation Modeling*, and *Applied Psychological Measurement*.



Chapter 3

Perspectives on Psychometrics: Interviews with 20 Past Psychometric Society Presidents

Abstract In this article, we present the findings of an oral history project on the past, present, and future of psychometrics, as obtained through structured interviews with twenty past Psychometric Society presidents. Perspectives on how psychometrics should be practiced vary strongly. Some presidents are psychology-oriented, whereas others have a more mathematical or statistical approach. The originally strong relationship between psychometrics and psychology has weakened, and contemporary psychometrics has become a diverse and multifaceted discipline. The presidents are confident psychometrics will continue to be relevant but believe psychometrics needs to become better at selling its strong points to relevant research areas. We recommend for psychometrics to cherish its plurality and make its goals and priorities explicit.

This article is submitted for publication as Wijsen, L. D. & Borsboom, D. (under review). Perspectives on psychometrics: Interviews with 20 past Psychometric Society Presidents. Submitted to *Psychometrika*

3.1 Introduction

In the history and sociology of science, the focus often lies on a chronology of important events, such as the making of significant discoveries and the emergence of scientific theories. When there is mention of the persona behind these discoveries, studies often stress the scientist's contributions. In the history of science, the emphasis thus often lies on specific contributions and written sources, such as research articles and dissertations. Examples of such studies on the history of psychometrics are Van der Heijden & Sijtsma (1996), Jones & Thissen (2007), Wijsen, Borsboom, Cabaço, & Heiser (2019), and Bennett & Von Davier (2017). However, scientists are not only tied to their discoveries or theories; they often entertain thoughts and visions about how science should operate, which cannot always be found in these written sources.[7] Due to their close involvement in a specific research area, it is likely that researchers have relevant ideas about historical, current, and future developments. This article presents exactly those thoughts and visions of researchers, psychometricians in our case, about the historical, current, and future directions of their field. This article thus sheds light on the first-person narratives that most historical studies usually overlook.

This project was inspired by the research methodology of oral history. Oral history is a branch within historical research that focuses on collecting personal testimonies of people who have witnessed a particular period or event (Abrams, 2010; Thompson, 2017). Oral history studies move the focus from written archival sources to the memories of people and invites these people to share their memories in interviews. In the history of science specifically, oral history invites scientists to share their memories of doing research and shed light on dilemmas and choices they encounter in their daily jobs. Such projects are becoming increasingly popular. Especially in the United States, many university libraries and research institutes now have access to large collections of interviews with scientists, predominantly from the physical sciences (Doel, 2003). Other examples of oral history projects in science are Wright & Ville (2017) in economic history, Baer, Jewell & Sigelman (1991) in political science and Smith & Rennie (2014) in evidence-based medicine.

With this oral history project, our aim was to provide a detailed and nuanced account of the history of psychometrics by asking prominent psychometricians to share their knowledge and memories of their personal career and the history of psychometrics. However, an oral history project also presents the opportunity to look at history that is actually in the making (Weiner, 1988). Most of our interviewees are still active in

[7] An example of a history of psychology in which the scholars and their views play a more central role is Dehue (1995), which gives an overview of the different debates in Dutch psychological science in the 20th century.

psychometric research or practice, and those that are not are often still involved or at least interested in current developments in psychometrics. So, not only do the interviews gain access to knowledge of psychometrics' history that otherwise might have gotten lost, the interviews enabled us to find out how psychometricians perceive current and future developments in psychometrics. Our oral history project thus investigates both the history of the field, as well as psychometricians' perspectives on current and future directions.

Groenen and Van der Ark (2006) describe the interviews they held with 12 prominent psychometricians with the purpose of investigating the current status of psychometrics. These interviews focused specifically on specific models and techniques that have either been influential historically speaking (such as Item Response Theory or Structural Equation Modeling) or interesting developments in contemporary or future psychometrics, such as data mining and Bayesian analysis. Though several of these developments are also mentioned by our interviewees in the interviews on several occasions, the focus of our analysis lies less on describing these concrete examples of models and research traditions, and more on the underlying motivations and reasons *why* psychometricians do research in a particular way. So, besides having a descriptive purpose, our paper also aims to analyze the given answers on a deeper level. For example: can we distinguish different types of approaches of doing psychometrics, and how do these approaches contradict each other? What are the different attitudes we find in relation to the future of the field, or with respect to other research areas? The qualitative analysis in this paper thus aims to uncover the different perspectives on psychometrics held by our interviewees.

One of the reasons it is particularly relevant to ask psychometricians to reflect on their own research domain is because psychometrics has a complicated position with regards to its close neighbors: psychology and statistics (Groenen & Van der Ark, 2006; Borsboom, 2006; Sijtsma, 2006). Psychometrics' origins may be nested in psychology, but its current course diverges in many directions (one of them being statistics), and this results in a multitude of approaches and perspectives on what psychometrics should offer. Should psychometricians affiliate more with the psychologists, and work on building psychological theory and explaining human behavior, or should they focus on designing statistical methods that are valuable to export to other fields as well? And what do they believe will happen to psychometrics in the following decades? Do psychometricians expect psychometrics to remain a successful research area in the future, or are there challenges ahead which psychometrics first needs to overcome?

In this project, we invited psychometricians to share their perspectives on such questions regarding the past, present, and future of psychometrics. The interviews provided a wealth of historical knowledge and interesting ideas that we cannot all incorporate in this article. For the sake of openness of data and preserving the richness of the interviews,

we decided to compile the revised transcripts in a book (Wijsen, forthcoming) so that the entire interviews will be accessible to people who are interested in reading the stories of our presidents. This article though is a more in-depth qualitative analysis of these interviews, and addresses several of the topics, themes, and dilemmas that are important to the presidents and the authors. Ultimately, we show how diversely psychometricians perceive their own field, and that psychometrics is not restricted to one approach only. In the discussion, we elaborate on how the interviews inspire a range of historical and philosophical questions for further research.

3.2 Methods

We invited 36 presidents of the Psychometric Society to participate as respondents in our project. The rationale for this choice lies in the fact that presidents of the Psychometric Society are key figures in psychometric research and are democratically chosen by the psychometric community; their reflections on psychometrics are therefore intrinsically interesting and worth preserving. We approached the presidents through a personal invitation, in which we asked the presidents to contribute to an oral history project about the history of psychometrics. Twenty-one presidents accepted our invitation; one president eventually canceled the appointment. A small possible source of selection bias is that our location, the Netherlands, and our attendance at the IMPS meeting in Asheville, North Carolina, made it relatively easy to interview people who reside in the Netherlands or who attended this conference. We also see that the older presidents were more inclined not to accept or respond to our invitation. Reasons for declining our interview were geographic location, old age, or not considering it an important cause. Some presidents did not respond to our invitation. The 20 interviewees who accepted our invitation were president of the Psychometric Society for a period of one year sometime between 1982 and 2013.

All interviews were held in person, either at people's homes, their work offices, in a public space, or at the International Meeting of the Psychometric Society (IMPS) in 2016 in Asheville, North Carolina. The interviews took place between April 2016 and October 2017. The interviews took between 45 minutes and one hour and were videotaped. The interviewee signed an informed consent in which he or she consented to use the material for this research project.

The questions of the interview were built around four topics: the respondents' professional career, their views on the relationship between psychometrics and other scientific disciplines, the history of psychometrics, and future directions of psychometrics. The questions were organized in a semi-structured interview format, which served as a general guideline. A subset of questions was posed to all candidates, but each interview allowed enough space and time to discuss topics that were interviewee-specific. The questions for the interview were sent to the interviewee beforehand if so requested.

The interviews were first transcribed in Inqscribe (Inquirium, 2013), and then roughly edited: the edited texts are as close as possible to the original transcriptions[8] but modified into readable and accurate English. When we were not completely certain about the exact wording used by the interviewee, we contacted the president and asked for rectification. The quotes from the interviews in this article are selected from these modified versions. For the sake of accuracy, the quotes were not taken from the more thoroughly revised versions that will be included in the compilation. After editing, we performed a qualitative analysis of the interviews: we identified the most prevalent themes, partly based on the themes already provided by the questions, partly based on the input by the presidents, and collected sections from the transcriptions for each individual theme. A selection of themes and corresponding quotes we thought were most relevant is discussed in the results below.

3.3 Themes

Improving our understanding of the history of psychometrics was the main reason for doing an oral history project, and is thus the starting point of this chapter. Before we continue with the presidents' perceptions, we will sketch a general (historical) framework that helps to contextualize the interviews.

Psychometrics originated at the end of the 19th century and early 20th century, with the work of academics like Francis Galton, Karl Pearson, Charles Spearman, and Louis L. Thurstone. Since, it has seen a number of shifts which closely resemble the four generations of test theory that Paul Holland (one of our interviewees) has conceptualized (described by Dorans, 2011). Holland's delineation starts in the early 20th century when test theory's first generation started with developments in classical test theory, reliability, and validity. The second generation, which started in the 1940s and peaked in the 1970s, was concerned with the development of models for item-level data. The third generation, which started in the 1970s, focused on the statistical advancement of item-level models. The fourth generation attempts to bridge the gap between the psychometrician and the testing enterprise, by developing methods for Differential Item Functioning or Test Equating.

When we transpose this delineation to psychometrics, we find that the four generations lack a clear role for factor analysis (a first-generation development and, as we will see later, considered crucial in the history of psychometrics) and Structural Equation Modeling and Multidimensional Scaling as part of the second and third generation. Moreover, we consider the fourth generation to be broader than just bridging the gap

[8] The modified transcriptions will become available in a separate publication (Wijsen, forthcoming); currently, transcriptions are available upon request.

between psychometrics and the testing enterprise: as we will see below, many fourth-generation psychometricians aim at finding connections with a variety of other sciences and enterprises, not just the testing industry.

Importantly, Holland argues that none of these generations have permanently ended: all generations – though some might have drastically shrunk over the years – are still active research domains, and *Psychometrika* still publishes research from these four domains. The most cited papers from the past two decades concern topics in Structural Equation Modeling, reliability estimates, and advances on a variety of latent variable models (a mixture of topics from different generations). Articles on Item Response Theory still make up a significant part of *Psychometrika's* content, and historically speaking, articles on the analysis of proximities have also been one of *Psychometrika's* pillars (Heiser et al., 2016). More recent directions are cognitive diagnosis, Bayesian methods for model estimation, and Computer Adaptive Testing. What is interesting about this list is that topics like the replication crisis, questionable research practices, and the practice of educational measurement – exceptions granted – are usually not addressed in *Psychometrika*. *Psychometrika* mainly publishes in-depth theoretical and technical papers, not commentaries on research or testing practices. Psychometrics, as understood in this paper, is thus a highly technical, abstract, and model-based research domain.

3.3.1 Key Moments in the History of Psychometrics

In the interviews, we asked the presidents how they perceive the history of psychometrics, and especially what they believe were psychometrics' key moments and main achievements.

One of the questions we asked is what the presidents believe is the most significant work or the most important psychometrician in the history of psychometrics. The most common answer (given by eight interviewees) was that this must be Lord and Novick's *Statistical Theories of Mental Test Scores* (1968). *Statistical Theories of Mental Test Scores* came out at ETS (Bennett & Von Davier, 2017) and was one of the first works in psychometrics to give an axiomatic treatment of classical test theory (Traub, 1997). Its publication took place in the midst of the shift from Classical Test Theory to Modern Test Theory, possibly the quintessential paradigm shift in psychometrics. Though classical test theory was strictly speaking never falsified, the latter became dominant in most psychometric research. Lord & Novick (1968) is one of the first comprehensive works to treat topics from both classical and modern test theory. Brian Junker praises it for having "everything from factor analysis to IRT and other things that are relevant to standard measurement questions in psychometrics. […] there is a real effort to connect psychometrics to current thinking in statistics." Ivo Molenaar praises it for being:

> on the transition of the old classical correlation-based and classical test theory-based models, to the item response models and latent trait models. […] Fred Lord was the classical one, and Mel Novick brought in the logistic models, which was definitely a very important step for the psychometric community as a whole.

This strong consensus on the central importance of Lord and Novick's *Statistical Theories of Mental Test Scores* is remarkable and invites further research on the effect the work has had on the development of the field. Some presidents go further back in time to the early 20th century and consider either Charles Spearman or Louis L. Thurstone, the founders of factor analysis, as the most important psychometrician in the history of psychometrics. Klaas Sijtsma regards Spearman as revolutionary:

> He actually combined psychological problems he was struggling with, with the development of statistical tools that he needed to tackle those problems, and in a way, he is the founding father of classical test theory and factor analysis, which is not a small accomplishment; it is incredible.

Paul De Boeck states that, between Charles Spearman and Louis Thurstone, he prefers the latter.

> He [Louis Thurstone] was doing factor analysis, but not just to measure. His paper was called 'Vectors of Mind', so he wanted to explain the human mind. He both had an interest in measurement, and an interest in understanding how the mind functions.

Larry Hubert commends Thurstone for training and educating so many prominent psychometricians, like Paul Horst and Ledyard Tucker. And it was also Thurstone whom David Thissen admires most: "Thurstone made everything. Thurstone made the discipline; he came from nowhere, received degrees in things like engineering, and created quantitative psychology; he created scaling, he changed factor analysis into multiple factor analysis. He started the Psychometric Society." Willem Heiser and Robert Mislevy consider Lee Cronbach as one of the most influential psychometricians in history. According to Heiser, Cronbach's paper on the reliability coefficient is one of his most significant contributions (Cronbach, 1951), due to its applicability to practical problems in research, not only in psychology but also in medical science or other fields where measurement plays a central role. Mislevy praises Cronbach for thinking critically about psychological measurement and the inferences or conclusions you can draw based on certain data, referring here to

generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972): "he laid down some real mileposts, about how psychometrics is not just about measurement, it is about the quality and the nature of inferences that you're making."

Some presidents do not mention specific people, but rather focus on a typical psychometric idea that was historically significant. For example, Peter Bentler mentions the theory of error as an essential scientific contribution by psychometrics:

> Very influential was the idea of errors in measurement, which of course, had been around for a long time in astronomy – it is not like Spearman invented it - but Spearman thought about it in a way that made it relevant to psychological measurement.

Jos ten Berge agrees: "The very simple fact that when you measure someone's intelligence twice, you don't get the same results, means that at least one of the two measurements cannot be correct, and that must be error." Not only is the idea of the quantification of error in measurement an important scientific contribution of psychometrics, but it also marks the attitude of the psychologist or psychometrician as a researcher. Jos ten Berge argues the following:

> It is a very interesting fact that psychologists have a routine of evaluating their measurements, for instance, by reliability and validity studies. It is a form of self-criticism that often isn't sufficiently appreciated. It is a very beautiful situation: a discipline that distrusts its own results.

The conceptualization of measurement error and its incorporation in psychometric models are thus seen as unique contributions of psychometrics to the sciences. Moreover, these contributions characterize how the psychometrician practices research: with a strong awareness of the imperfection of (psychological) measurement. Ten Berge's remark underscores that the characteristic viewpoint of the psychometrician involves the recognition and appreciation of the problems involving psychological and educational measurement.

3.3.2 The Dark Ages of Psychometrics

According to several presidents, psychometrics' most important contribution to society is psychological and educational testing. Testing has pervaded several phases in people's lives, and psychometricians turned it into a standardized and reliable enterprise. However, measurement and testing do not only resonate in the ears of some of our respondents as something that is only positive and for a good reason. Despite the fact that the controversial

part of the history of psychometrics was not an official interview topic, some presidents bring it up themselves, often torn between psychometrics' controversial history on the one hand and its important achievements on the other. When David Thissen states that it was indeed testing that put psychometrics on the map; twice, he states that this was "for better or for worse." Jacqueline Meulman says that she:

> was amazed by how many bad things had happened in psychometrics, I was flabbergasted. On the other hand, I was intrigued by the mathematical background of the methods I was reading about [...]. Although I did realize that many of the great psychometricians didn't have very good political backgrounds, I was intrigued by the methods themselves [...].

The interviewees refer here to the controversial history of mental measurement, which was strongly intertwined with 19th and 20th-century politics, and especially eugenics. Eugenics – a scientific and political movement that aimed to improve the genetic quality of the human population, which thrived late 19th and early 20th century (Buss, 1976; Chitty, 2007; Norrgard, 2008) – was a popular ideology among many psychometricians, among which Charles Spearman, Lewis Terman, and James McKeen Cattell. In these times, the measurement of intelligence was often misinterpreted and misused to attribute differences in intelligence test scores to genetics (Richards, 1997; Jackson & Weidman, 2004). Predominantly during the late 19th century and early 20th century (though not exclusively so), differences in intelligence served as 'scientific' proof for the claim that some groups (Afro-Americans, women, people of lower classes) were less intelligent and thus less worthy than upper-class white males. And though the Psychometric Society did not have an explicit eugenic agenda (or any political motivation for that matter), at least one president entertained similar ideas. Henry Garrett, president in 1943, supported the idea of hereditary racial differences in intelligence and racial segregation (Winston, 1998). The history of psychometrics is thus not a sequence of one groundbreaking scientific achievement after the other, nor were all psychometricians always distrusting of their results.

Other presidents also refer to the adverse effects of psychometric research. Bill Stout states that when done well, psychometrics can be very important, but psychometricians have also sometimes "oversimplified a very complicated subject." Here, Stout refers to the Bell Curve controversy (Herrnstein & Murray, 1994), a more recent example of how differences in intelligence scores are used to justify differences between races and social groups. Larry Hubert is highly critical of psychometrics' past, and where other presidents see testing as a relatively positive contribution of psychometrics, Hubert is not so sure: "[…] I'm not sure if all in all the idea of measuring intelligence hasn't brought more ill

stuff than it has brought good stuff. The whole politics of race and psychometrics is not a very happy one." Though the dark ages of psychometrics were not an official interview topic, several presidents touch upon them on their own initiative, implying that these dark ages should not be overlooked in further historical research.

3.3.3 The Relationship between Psychometrics, Psychology, and Statistics

As we discussed in the introduction, what is intriguing about psychometrics is its position relative to other disciplines. Though psychometrics originated in psychology, it is now closely affiliated to statistics as well. In this section, we will discuss how the presidents perceive the relationship between psychometrics and two of its closest neighbors: psychology and statistics.

Psychometrics, Psychology and Educational Measurement

The relationship between psychometrics and psychology is hard to define, but the detachment between psychometrics and psychology (and also the detachment between psychometrics and educational measurement) rises to the surface in several interviews. What the psychometricians disagree on is whether this detachment is indeed an issue, and in case it is, how psychometricians should act on it.

A particularly vivid illustration of the disconnected relationship between psychology and psychometrics is formed by the similarly detached attitude of some of the interviewees towards psychology. Some presidents express a certain ignorance of or lack of interest in what is going on in psychological research: they explicitly mention knowing little of psychology, or just not being interested in it. For example, statistician Bill Stout stresses the importance of statistics in psychological research but mentions not knowing enough about what is going on in the field of psychology to see how psychometrics can contribute. Jacqueline Meulman expresses her discomfort with topics in psychology or educational measurement and states she feels more at home in biostatistics. Though appreciative of fellow psychometricians doing psychological research, their own interests lie somewhere else.

This indicates an important change with respect to the early 20th century because it is hard to imagine a similar approach to psychology and psychometrics in the early days of psychometrics when psychometrics and psychology were still in a close relationship. The remarks of some of the presidents show that it is currently possible to be a successful psychometrician and a president of the Psychometric Society, without having either a background or an active interest in psychology. Being successful in psychometrics and being a president of the Psychometric Society, therefore, does not require a strong connection to psychology or educational measurement: having strong ties with mathematics or biostatistics is equally relevant and appropriate. Modern psychometrics has thus evolved

into a field that is no longer dedicated to psychology alone and can no longer be defined as psychology's statistical counterpart; instead, psychometrics has developed ties with different fields, which shows in the backgrounds and interests of the presidents of the Psychometric Society.

Several presidents argue that standardized testing or educational measurement is the most important contribution of psychometrics. However, some stress that psychometrics also has trouble reaching educational measurement: similar to psychology, educational measurement is missing out on some of the newest psychometric methods. Susan Embretson explains that this is because "testing is the hardest thing to change"; people in education are slow in adopting cognitive theory for item construction. According to Jacqueline Meulman, educational measurement is missing out on psychometrics because "major testing institutes in the US don't use the work of psychometricians, and there are even institutes or agencies that do testing that use nothing that comes of out of the psychometric community." However, the detachment might be less severe than with psychology: psychometricians like Wim van der Linden and Hua-Hua Chang also see many possibilities for psychometrics in educational measurement, especially for adaptive testing. According to Van der Linden and Chang, there is high demand for adaptive methods and they see this continuing in the future.

There are a number of possible explanations for the growing distance between psychology and psychometrics. David Thissen explains that, before the 1950s, a psychologist was also trained in psychometrics, but for the sake of the grant system, psychology departments are divided into subfields. "It is now almost inconceivable to get to this state of the art in more than on one of these subareas, in one brain. You can never know enough." In other words, one becomes a social psychologist, a developmental psychologist, or a psychometrician, and there is very little mingling between the three professions. Related to this, Jan de Leeuw states that he also finds it the job of the psychologist, not of the psychometrician, to engage with building psychological theories. According to De Leeuw, the psychologist and the psychometrician simply have different job descriptions, which means that the work they are doing is fundamentally different.

A second explanation for the growing distance between psychology and psychometrics has to do with how psychometric research is communicated to external parties. Bengt Muthen, Larry Hubert, and Peter Bentler express their opinion that *Psychometrika* or other psychometric literature can sometimes be too narrow in terms of content, and perhaps also too technical and too theoretical for the psychologist or educational researcher to read and use. Consequently, *Psychometrika* has become out of reach for applied researchers without thorough psychometric or statistical training. Psychometrics might thus have become too much of a niche, and consequently, detached from psychology.

Psychology first!

For several presidents, the growing distance between psychology and psychometrics is a point of concern. Klaas Sijtsma states that he now encourages "everybody to engage in theory building. So, to become a psychologist, rather than a psychometrician." He pleads for a more unified psychology, where once again people are trained both as a psychometrician and psychologist. De Boeck also pleads against using psychometrics as purely a statistical toolkit: "I think psychometrics is a way of thinking about substantive issues, and it's possible to come up with ideas, substantive ideas, based on a certain way of understanding psychometric models." According to these presidents, psychometrics is not just a toolbox of purely statistical, data-analytic models, but a set of models and techniques that can inspire substantive thinking about psychological problems and thereby aid psychology theory building.

A reason why building psychological theory is no longer one of psychometrics' priorities is given by Susan Embretson:

> There is a whole breed of psychometricians out there who seem to have less of a substantive background, and I do not think that's a good thing. I think they might be dealing with rather narrow statistical issues that are not really going to make a difference in the discipline […]. So, I really see a necessity to keep quantitative methods attached to a discipline so it can influence that discipline.

According to Embretson, psychometricians can sometimes be too involved with technical details, whereas they should pay more attention to what they can contribute to psychological research. As mentioned earlier, *Psychometrika* mostly publishes articles on narrow, statistical issues, rather than articles that are relevant and readable for the psychologist. Psychologists might, therefore, not be inclined to look for relevant literature there.

However, the reason for the detachment does not only lie in psychometrics' court. Several presidents mention the lack of interest of the psychologist in applying proper psychometrics. When we ask James Ramsay to identify the relationship between psychology and psychometricians, he answers:

> I would say it is both distant and uneasy because the psychologist needs psychometricians badly, but quite frankly, once they have what they need, they do not want to hear anything else, so statistically speaking, it is a very conservative community.

It is hard to escape a sense of disappointment or frustration here. Psychometricians are not able to get their expertise across, whereas helping psychologists with their methodological problems is often considered part of the job description of the psychometrician. The psychometrician is supposedly the consultant who offers statistical or methodological advice, but psychometricians can only do their job if the psychologist seeks the psychometrician's help when in need. In practice, this does not happen frequently enough, and that is considered a shame. Wim van der Linden states that psychometricians "could be a major support to psychology, make their measurement rigorous, and then plan their experiments better, help them model. […] it could feed psychology." Psychometrics could thus provide valuable input for the psychologist, which the psychologist is now missing out on.

The interviews show that the relationship between psychology and psychometrics is nothing short of complicated. What makes the psychology-psychometrics relationship even more challenging, is that psychometrics is also strongly affiliated with statistics, the topic of the next section.

Psychometrics and statistics

After psychology, statistics is probably psychometrics' closest kinship, and the relationship between the two was frequently touched upon in the interviews. According to Brian Junker, the separation of psychometrics and psychology is not necessarily a reason to worry: "In a certain sense, psychometrics is by definition tied to psychology, but the methods are really just the methods of latent variable modeling for individual differences, and that may or may not be tied to psychology." According to Junker, psychometrics may have its origins in psychology, but this does not imply that psychology should be its only connection. Many presidents stress that it would be beneficial for psychometrics if it were to extend its influence to other fields. They believe psychometrics should make more effort to be taken seriously by other fields, like statistics, since it could make important contributions there as well.

Willem Heiser uses the metaphor of a river system to describe the relationship between statistics and other disciplines with a strong quantitative component:

> A river system starts with small little rivers, and which is where I consider the various disciplines, like biology, psychology, economy, econometrics, chemistry. Those are the areas where people do quantitative things. Sometimes, they invent something for themselves which is useful for others, and then these techniques that are invented in a substantive area go down the stream to the big river. The big river is statistics, so to speak. That is where everything ends up.

According to Heiser, scientific disciplines with a quantitative focus each develop their own statistical methods, which at first are devoted to solving a specific substantive research question, but then get stripped from substantive interpretation. These models are subsequently free to move from the small river to the big river of statistics, which is filled with models developed in a wide variety of research areas. Not uncommonly, quantitative methods developed in one river find their way to other disciplines as well. An example of such a method in psychometrics would be factor analysis, which was originally developed to describe general intelligence, and has now found its way to other research areas both in and outside psychology (Young & Pearce, 2013).

The close connection between statistics and psychometrics becomes clear when we find that a number of presidents do not have a background in psychology, but in statistics or mathematics. Paul Holland articulates this close connection between the two: "I think that psychometrics has a very strong statistical side, I keep thinking of psychometrics as being part of statistics, not so much 'psycho'. Even though the guys that invented the field all came from psychology." Like Willem Heiser, Paul Holland stresses that methods developed in psychometrics are no longer restricted to psychological research alone, and can be used by other disciplines. Taking Holland's perspective a bit further, we might say that psychometrics has lost its 'psycho'-affiliation throughout the years, and became a type of modeling that is relevant for a variety of research domains (psychology, sociology, medical science, artificial intelligence), and can be gathered under the statistics umbrella.

Even though psychometrics and statistics have a close relationship, several presidents point out that psychometrics has a problem making that connection beneficial for both sides: there is plenty of proper, technically well thought out psychometric work that is useful for the statistician but is not recognized as such by other statisticians. Jan de Leeuw gives a reason why original psychometrics did not strike a chord with the statisticians: "It was mostly because of the way the original factor analysts, who were psychologists, like Spearman and Cattell, presented [factor analysis] as some magical tool that could discover laws of nature by simple inductive data analysis." Interestingly, the same magic-jargon is mentioned by Bengt Muthen, who says that "statisticians think of that [factor analysis and structural equation modeling] as hocus pocus machinations." Psychometricians magically pulling 'factors', such as intelligence, out of the hat did not sit well with the statisticians, who were most likely less interested in making strong substantive claims about the identity of latent variables than the psychometricians and psychologists at the time.

Moreover, some interviewees point out that on a number of occasions, research that was being done under the name of statistics, had actually already been done before in psychometrics. But because psychometrics is too much of a niche field, researchers from other fields simply do not know it had already been done before. And this leads to

frustration among some of the presidents since psychometrics could, in fact, contribute a lot to the field of statistics. According to Muthén:

> [...] it is a strong tendency in statistical journals to refer to early statistical articles referring to the psychometric literature [instead of referring directly to the original psychometric literature] [...]. It seems psychometric publishing seems to be too separated from general mainstream statistical modeling […].

Interestingly, the public relations issues of psychometrics seem to come up both with the psychology-oriented presidents and with the statistics-oriented presidents: psychometricians are not able to reach out to either group and fail to receive acknowledgment for their work.

3.3.4 The Identity of the Psychometrician: A Multitude of Approaches

The sections above show there are multiple ways how the psychometricians perceive their own field, and that contemporary psychometrics consists of a variety of approaches, each with their own ideas and visions. Below, we distinguish between five approaches we have recognized in the interviews. Our intention here is not to categorize each respondent and define them as a specific type of researcher, but to show there are different ways in which psychometrics research can or should be practiced, each prioritizing different characteristics or elements of psychometric research. The types discussed below underscore the plurality of approaches in a field that, to the outside, might seem relatively uniform.

The Psychologist

First of all, unsurprisingly perhaps, we identify the psychometricians who identify themselves as both a psychometrician and a psychologist. The psychology-oriented psychometrician uses psychometrics as a way to improve psychological understanding and always has a substantive interest. According to the psychology-oriented perspective, psychometric models do not only describe or summarize psychological data but can help in understanding or explaining the data as well. The division between the psychometrician and the psychologist then becomes rather fuzzy: psychometricians who are driven by substantive questions take on a double identity (being both a psychometrician and a psychologist) rather than identifying themselves as solely a psychometrician. For reasons cited earlier, people like Klaas Sijtsma, Susan Embretson, and Paul De Boeck are psychometricians who have a psychology-oriented approach.

The Consultant

Closely related to the psychology-oriented approach, but not entirely equivalent, is the consultant approach. The consultant aims to maintain a close relationship with psychologists and encourages collaborations, in which the psychologist comes up with a substantive research question, and the psychometrician offers methodological advice. The difference between the psychologist approach and the consultant approach is that the psychometricians of the first kind have an intrinsic interest in psychological theory and uses psychometrics as a way to build psychological theories, whereas the psychometrician with a consultant approach prefers to aid psychologists in solving methodological and statistical problems and leave the actual theory building to the psychologist. Peter Bentler and Bengt Muthén, who often collaborated with psychologists or other applied researchers and helped them solve complex methodological problems, might recognize themselves as taking up such a role in their research.

The Data-Analyst

Third, we find that a number of presidents have more of a data analytic approach. These psychometricians view psychometrics as a toolbox that contains a set of models that are mostly of the latent variable type, which they consider applicable to a wide variety of data and disciplines. Though some of these models were perhaps originally designed for psychological measurement, in a data-analytic approach, these models are not necessarily used as substantive models and can be translated to several types of data for different types of purposes. The goals for the data analyst are usually not explaining the data or understanding the underlying mechanisms (which would be major motives for the psychology-oriented psychometrician) but rather to make predictions, or summarize the data. Brian Junker, as quoted earlier, considers psychometric models to be translatable to all sorts of research problems. This view aligns with the data-analytic approach.

The Engineer

A fourth type we encountered is the engineer. Engineers are people who are interested in 'making' technologically advanced artifacts, which then find a clear application in society. Examples of such artifacts in psychometrics are innovative types of tests, like computer adaptive tests or simulation assessments, but also software programs. These applications then find their way to testing agencies, educational measurement, or the scientific community. Through these artifacts, the engineer may try to explain human behavior or solve challenging technical problems, but this takes place through a real-world application, rather than doing foundational or theoretical work only. People like Hua-Hua Chang, Wim van der Linden, and Robert Mislevy are co-builders of such applications and share an engineering-approach.

The Mathematician

Lastly, we distinguish the mathematician who gains most joy out of proving a mathematical theorem or solving a technical problem, without necessarily feeling the need to find an application or answering a substantive research question. The mathematician approach does therefore not require collaboration with psychologists or other applied researchers. For the mathematician, knowledge for the sake of knowledge (not for the sake of application) is sufficient. Moreover, the indisputable quality of mathematics – proving a theorem for once and for all – has an incredible appeal to some of the presidents. Jos ten Berge stresses that what he likes so much about psychometrics is "the absolute certainty with which you can decide about what is true or isn't true. The mathematical part of it." This sentiment is also shared with Jan de Leeuw, who finds psychology too "debatable, or uncertain, or up in the air", and who appreciates the beauty of mathematics.

Two Dimensions of Psychometric Research

Naturally, a psychometrician does not necessarily fall under only one of the categories above: a combination of approaches is equally plausible. For example, someone who is a designer of technologically advanced tests – whom we might characterize as having an engineering approach –, may also be interested in learning mechanisms in school children and thus have a substantive or psychological interest as well. For this reason, we propose to summarize these categories in two dimensions, one ranging from 'psychology' to 'statistics', the other ranging from 'theoretical' to 'applied'. Our respondents differ from each other in whether their research is driven by psychological questions or technical statistical issues, and at the same time, they differ in how strongly they concern themselves with applied or theoretical topics. Someone with a mathematical approach is more on the theoretical and statistical side of these dimensions, whereas the psychometrician with a strong psychological interest can be located in the psychology/theoretical corner (or more on the applied side, if this psychometrician has a strong focus on doing empirical research). These dimensions thus describe core aspects of the multifaceted identity of psychometric research.

3.3.5 The Future of Psychometrics

The interviews provided an excellent opportunity to invite the presidents to take a look into the future of psychometrics and ponder on possible directions psychometrics might take. Some presidents think psychometrics will continue to remain relevant. Jos ten Berge stresses that, since psychologists do not have the technical training that psychometricians have, there will always be a need for psychometricians. According to David Thissen: "[...] testing will continue to develop and continue to be a thing that is done for placement in education, in jobs. […] I think testing still has some decades, if not centuries in it."

Testing thus remains an important application of psychometrics. Analyzing test data well and making the right decisions based on test scores are still crucial in today's society, and will most likely continue to remain crucial in the upcoming decades. Moreover, testing now transcends traditional paper-pencil formats, and new types of tests are continuously being developed. The expertise of the psychometrician is therefore crucial and relevant and will remain so in the future.

However, the future relevance of psychometrics does not seem guaranteed. A number of interviewees express a certain sense of uncertainty with regards to a fruitful future for psychometrics. Though the interviewees disagree on what they believe the future holds, several presidents agree that a prosperous future for psychometrics is not a given. Psychometricians will have to put in the effort to make themselves relevant.

Some presidents point out that psychometrics has a serious PR problem and has to work hard to be heard, whether it is by psychologists or by other possible collaborators, and many see challenges in selling psychometric research to relevant parties. In fact, Wim van der Linden considers the inability of psychometrics to market itself as psychometrics' biggest pitfall. He blames this inability on the slow development in psychometrics of making good user-friendly software, which would have paved the way for selling psychometric models at an earlier stage. Robert Mislevy states that "it is easier to get people to recognize the value and the use of psychometric techniques if you do not call them psychometric techniques until you have worked with them for a couple of months at least!" Even though the presidents think it is crucial that psychometric knowledge is not lost to the test of time, psychometrics will have to make up a plan to remain influential. Mislevy continues: "there are very rapid advances today in technology, in psychology, in learning analytics, and the biggest challenge of psychometrics is not getting left in the dust."

When asked about what the future holds for psychometrics, some respondents refer to the big data era, and how psychometrics could contribute to such new developments. Some say that the big data era provides an opportunity for psychometrics, and that again, we should not miss the boat. Ulf Böckenholt is full of optimism: "We live in the age of big data, the age of self-quantification. I carry a Fitbit. It is the dream of the psychometrician!" And, according to Paul Holland, "The future of psychometrics is about the open-mindedness of all the different varieties of the ways that people collect data and try to draw conclusions and to make sense of it." It is the age of big data, and human response data is anything but extinct. In fact, more and more different types of data, in need of thorough analysis, are coming our way. And, according to Hua-Hua Chang, psychometricians have relevant knowledge that other researchers do not:

> Everyone is talking about big data, but what is big data? How is the data collected? I think our psychometricians should do a good job of making sure data is collected reliably. How was the data collection designed? Does it have high validity? [...] That will make psychometricians even more important.

Thus, big data need to be analyzed appropriately, and psychometricians have the tools to get involved, also when the nature of these data is significantly different from traditional testing data.

But even though the big data movement seems more than promising, Jacqueline Meulman warns for the hype. According to Meulman, both psychometricians and statisticians should be critical of this development. Instead, psychometricians should claim back their own field:

> They should say, 'psychometrics is our area, and testing is from our origins, and we should claim it back.' I am amazed sometimes by things I see on the Internet, that major agencies that do testing have no clue what psychometrics is all about.

Meulman stresses that it is by no means her intention to ignore developments that are going on in data science, but that it is essential to be on guard with these modern trends, and also to remain influential where psychometrics has always been needed the most: the testing industry. Ivo Molenaar also warns for the rise of big data:

> I think that they [the psychometricians] have more computational possibilities now and have what they call big data [...]. I am getting old-fashioned, so I think maybe you should not collect that many data because it is only going to cause you problems.

Molenaar refers here to the danger of overfitting and the lack of critical thinking in a mostly computer-driven process.

The future of psychometrics is thus regarded with careful optimism. Several presidents believe that psychometrics will remain relevant for psychology and the testing industry. However, where some presidents stress the importance of opening up to contemporary scientific ideas, others explicitly warn for these new developments. Both sides are afraid psychometrics might remain too isolated and out of touch with the scientific playground.

What is worth noting is that, even though the interviewees were critical and sometimes weary of a fruitful future for psychometrics, they did not foresee the specific challenge psychometrics is facing in the spring of 2020. Major universities in California

are currently rejecting college admission tests from established psychometric agencies like SAT and ACT, because these tests would counteract diversity. Though psychometrics has a long tradition of research on test fairness and item bias, and these agencies have a long tradition of making admission tests as fair as possible, testing is still considered disadvantageous for poor, Hispanic, and Afro-American students. The ideology of meritocracy, so engrained in psychometrics and university policy, is thus shifting in favor of diversity and inclusivity. The next few years might prove to be even more challenging for psychometrics than the presidents had initially anticipated.

3.3.6 Recommendations

Psychometrics might thus benefit from a change of course. But what change? It is challenging to extract a single recommendation from all twenty transcripts. What we can safely conclude, is that contemporary psychometrics is essentially a pluralist research area, and it is this plurality that needs cherishing. This does not mean that we should just 'let things be pluralist' and each go our own way, which is perhaps what is happening now. Instead, psychometrics needs to make explicit what a plurality of goals and approaches actually entails. What are the avenues that psychometrics aims to tread? What is psychometrics' mission, and what are its priorities? Where and how does psychometrics want to contribute? We would recommend the Psychometric Society and other psychometric institutes to list their priorities and make a resulting mission statement public. Based on the interviews, these priorities could include: 1) building psychological theory, 2) improve educational measurement in terms of fairness or reliability, 3) construct and distribute user-friendly software for the analysis of behavioral data, 4) develop new methods for data analysis, and perhaps even 5) work on psychometrics' credibility. Not only does such a list of priorities make it easier to communicate to external parties what it is that psychometrics does and values (the lack of communication being a point of concern for several presidents), it can also offer guidance on relevant topics for sessions at meetings and the publication of articles. With this recommendation, we have no intention of preventing researchers from pursuing a path that is not listed as a priority. However, a more active policy may provide some clarity and guidance for a field that, if current trends continue, with time will only become more and more fragmented and diverse.

A second recommendation has to do with psychometrics' relationship with its past and how its history also shapes contemporary psychometrics. Early psychometricians like Francis Galton, Lewis Terman, and James McKeen Cattell were often devoted to a specific social ideal – often associated with the highly controversial ideas of eugenics – and they expressed these ideals in their academic work. It is interesting to see that contemporary psychometricians do not often engage in public debate – even when educational measurement is again part of a heated discussion – and *Psychometrika* rarely publishes

articles about such themes. Perhaps, psychometrics' controversial history functions as a warning against a strong social involvement. Instead, contemporary psychometrics engages in highly technical work that, on the face of it, often seems to be detached from social reality. Psychometricians' shyness for public expression does not help in improving their visibility, and importantly, it might lead to outcomes that are completely undesirable to the psychometrician (e.g., the possible decline of reliable measurement in schools or the rise of irresponsible data analysis). Whatever the reason for psychometrics' current absence from public debate, we would recommend psychometricians to engage in matters that touch upon their expertise, not only as a way to increase their visibility, but more importantly, because they have expertise that matters and it is their duty to share this with the world.

3.4 Conclusion & Discussion

First and foremost, the interviews testified to the fact that psychometrics is a multifaceted discipline, which creates tensions that are intrinsic to its organization: psychometrics is structurally related to different fields, and our interviewees disagree on which of these affiliations should be leading. Clearly, modern-day psychometrics has evolved into a much more diverse, but also more fragmented field than it used to be in the early 20th century when psychology was psychometrics' main focus area. Though most psychometricians agree psychometrics may have a successful future, they also express their concern about psychometrics not being able to reach out to other relevant areas.

The diversity of the field, both in current practice and perspectives on future directions, raises the question of how this diversity originates. One explanation for the diversity in psychometric research is that the Psychometric Society itself does not conduct a clear policy of where psychometrics should be heading. More than anything else, it is expertise and intellectual contributions that help decide whether someone becomes president of the Psychometric Society. Having a particular vision about the future of the field is not a requirement for the presidency. And even if a president is determined to adopt a specific policy in order to promote a particular approach of psychometric research, one year of presidency is often too short a period to leave a lasting impression on the psychometric research climate. The Psychometric Society, therefore, does not have one clear direction apart from promoting psychometric research in general and thus leaves plenty of room for a variety of approaches.

The diversity and fragmentation of psychometrics are not intrinsically problematic: the field is now reaching beyond its traditional boundaries, resulting in a wide variety of psychometric research projects all over the world, and that can certainly be considered a positive development for the field. But while the research topics and the psychometricians themselves have become more diverse and have only increased in number over time,

psychometrics is clearly having difficulty with connecting to both its home base, psychology, and other relevant fields. If psychologists, statisticians, or other applied researchers do not seek out psychometric expertise when in need, psychometrics is in danger of becoming a field that is only practiced within its ivory tower, isolated from other research fields or social applications. For psychometricians with a theoretical or mathematical approach, this may not seem very problematic, but most psychometricians wish to contribute to some area of application, and for them, a further detachment between psychometrics and other scientific disciplines is not a desirable prospect. In this regard, psychometricians will have to find ways to improve communication with researchers from other areas, and perhaps the Psychometric Society and other organizations can play a more explicit role here (for instance by making psychometrics journals more accessible to applied researchers or inviting a diverse group of speakers to speak at conferences). As we discussed in the section on recommendations, we would encourage the Psychometric Society to embrace psychometrics' plurality, and make explicit which goals psychometrics strives for, as to provide guidance in times of fragmentation. This is, of course, easier said than done, but for the sake of psychometrics' future, it might be of vital importance.

Though this study does not have philosophical aspirations, the interviews and our analysis generate several questions about psychometrics that might be relevant for future studies into the history and philosophy of psychometrics. Earlier, we briefly alluded to the shift from classical test theory to modern test theory, which took place in the 1960s and 1970s, which was accompanied by the publication of Lord & Novick (1968). The first question that arises immediately is whether the transition from classical test theory and modern test theory can indeed be understood in Kuhnian terms. Is there indeed a drastic shift in the meaning of terminology from one paradigm to another, or do these paradigms co-exist? Though Kuhnian or even Popperian readings of the history of psychometrics would indeed be valuable, our findings particularly invite a pragmatist reading of psychometric research. As we have shown in this paper, some psychometricians do not prioritize theory building, but rather the practical or predictive (pragmatic) value of a model or method. What it means for a psychometric model or method to be practical and how psychometrics is informed by a pragmatist philosophy are questions yet unanswered. A pragmatist reading of psychometrics would be particularly relevant since one of the inventors of pragmatism, William James (1907), is also one of the founding fathers of American psychology and psychometrics (Wijsen et al., 2019). His influence on the field, both in a historical and in a philosophical sense, has so far not received the attention it deserves.

Another topic worthy of further investigation is how psychometrics has become so diverse over time, and if it is representative of the diversification or specialization process in other fields. Can the patterns we found in psychometrics – an influx of researchers

with a different educational background, less familiarity with traditional research topics, an increasing detachment between people with substantive interests and people who are more technology and statistics oriented, a more diverse playground of research topics – be generalized to other scientific domains? Is this a pattern that is almost typical for advanced discipline formation, or is psychometrics unique in that sense? All in all, though the transcripts themselves do not formulate answers to the questions stated above, they inspire further research in these areas, which we would certainly recommend.

3.4.1 Some limitations

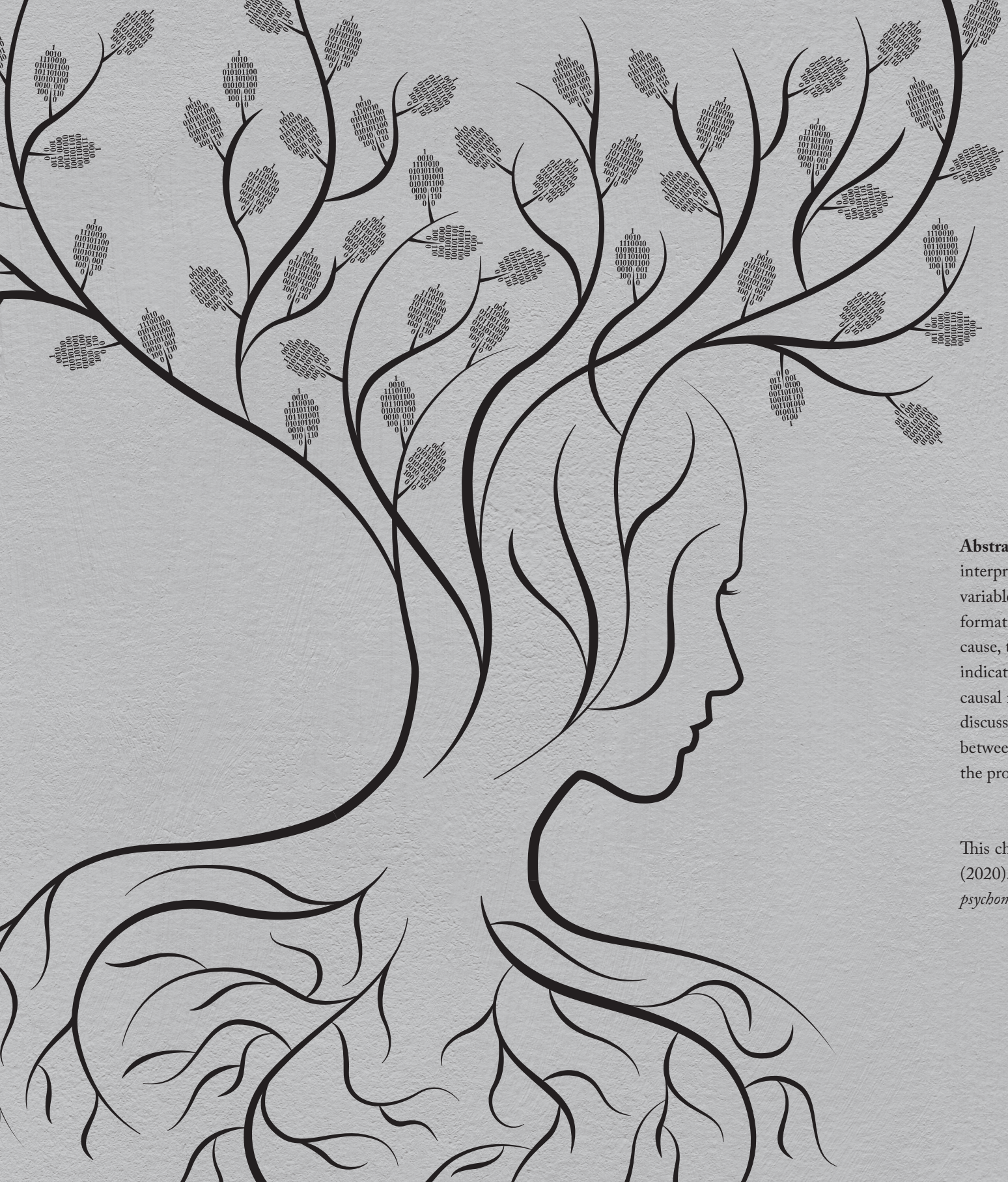
An important limitation of this project is that we only invited presidents of the Psychometric Society as interviewees, who are well known in their field and have already received credit for their work, rather than psychometricians who are perhaps lesser-known and are still building a career for themselves. An oral history methodology would indeed recommend inviting people who have not already received plenty of attention for an interview since their voices are less frequently heard than the voices of people who stand at the forefront. Their ideas might be substantially different from those of our presidents. This is, of course, a valid point, and perhaps a future project could focus on reflections of psychometricians who are at the beginning of their careers. There is however, a rationale behind choosing the presidents as our respondents. The presidents of the Psychometric Society have formed an integral part of 20th and 21st-century psychometrics. Their ideas and views represent – though not exhaustively of course – historical and contemporary developments in psychometrics and are therefore intrinsically interesting and relevant.

A second limitation concerns the possibility that the interviewees did not want to be too explicit about certain topics, knowing the interview was to be videotaped and used for research. For the sake of avoiding unpleasantries, they may have avoided voicing unpopular opinions or pointing fingers. However, even if this was the case, we find plenty of explicit ideas and visions in the transcripts, giving us reason to believe that the presidents were more than willing to share their ideas about and memories of doing psychometric research.

Retrospectively, we find some questions and topics are left unaddressed or received too little attention in the interviews. Most notably, we find that the interviews lacked focus on the historical adverse effects of psychometrics on scientific research and society. Though briefly addressed in section 3.3.2, we find that this highly relevant topic is deserving of a more thorough and critical investigation, and we recommend taking this up in further studies.

3.4.2 Acknowledgments

We want to thank Bengt Muthen, Bill Stout, Brian Junker, David Thissen, Hua-Hua Chang, Ivo Molenaar, Jacqueline Meulman, James Ramsey, Jan de Leeuw, Jos ten Berge, Klaas Sijtsma, Lawrence Hubert, Paul De Boeck, Paul Holland, Peter Bentler, Robert Mislevy, Susan Embretson, Ulf Böckenholt, Willem Heiser, and Wim van der Linden for participating in this project.



Chapter 4

Reflective, Formative, and Network Models: Causal Interpretations of Three Different Psychometric Models

Abstract This chapter gives an overview of three psychometric models, that can be interpreted causally. In the reflective model, a type of latent variable model, the latent variable is modeled as the explanation for correlations between the indicator variables. In formative models, these relations are reversed, as now the indicator variables construct, or cause, the latent variable. In the formative model, we make a distinction between causal indicator models and composite models. We also discuss network models, in which the causal relations are drawn between the indicator variables only. This chapter ends with a discussion of a number of controversies regarding these models, namely the relationship between data models and causal models, the causal status of individual differences, and the problems of generalization and interpretational confounding.

This chapter is adapted from Wijsen, L. D., Van Bork, R., Haig, B. D., & Borsboom, D. (2020). *Reflective, formative, and network models: Causal interpretations of three different psychometric models*. [Book chapter in press].

4.1 Introduction

A central topic in psychometrics is the construction of statistical models which can serve to connect theoretical constructs (e.g., intelligence, depression) to observable indicators of these constructs (e.g., IQ scores, DSM-5 symptomatology). These models, most of which are known as latent variable models, can be understood as causal models, in which the observable indicators relate causally to the theoretical entities in question. The theoretical construct is then considered an explanation for the responses on the observed variables. In other words: one person has a higher score than someone else *because of* his or her level on the theoretical construct. The most famous account of such a model is Spearman's two-factor theory of intelligence (1904), in which the construct of general intelligence was considered to explain people's test scores. In this chapter, we give an overview of three different types of causal accounts of psychometric models and explain some of the core assumptions for each of these models. We will also compare the models on a number of aspects and describe the situations in which these models properly apply.

The first account we discuss is based on the latent variable model (Spearman, 1904; Rasch, 1960; Lord and Novick, 1968; Mellenbergh, 1994b). In this account, it is assumed that the causes of human behaviors, which serve to produce observed variables in data analysis (e.g., answering an IQ-item, responding to an interview question), are *latent*, or unobserved, psychological entities. One gains access to these constructs by measuring them with the appropriate tests. Latent variable models provide a method to measure and theorize about such psychological constructs, and have come to occupy a prominent position in research in personality (McCrae and Costa, 1987; 2008), intelligence (Jensen, 1999) and psychopathology (Krueger, 1999; Caspi et al., 2014). Important latent variable models include the common factor model (Spearman, 1904), the Rasch model (Rasch, 1960), and the latent class model (Lazarsfeld & Henry, 1968). Mellenbergh (1994b) gives an overview of models that fall under this general statistical framework.

In psychometric applications, the latent variable model is often interpreted as a *reflective measurement model* (Edwards & Bagozzi, 2000), meaning that the variance in the observed scores on the indicators is hypothesized to be a direct reflection of variance in the latent variable, which serves to cause the variance in the indicators. In this chapter, we contrast this reflective measurement model with the *formative model*. In the formative model, the observed scores are not a reflection of the latent variable, but rather the cause of variance in the latent variable (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). As such, the latent variable is not the common cause of the indicators but an effect of the indicators. This means that, in contrast with the reflective model, the causal arrows run from the test scores to the latent variable.

The third model we will discuss is the *network model*, which has recently attracted an increasing amount of attention in psychology. The network model offers an alternative

method for thinking about the relationship between constructs and indicators in psychology (Borsboom & Cramer, 2013; Cramer, Waldorp, Van der Maas, & Borsboom, 2010). Rather than modeling the relationship between latent variables and indicator variables, the network approach focuses on modeling the relations between indicator variables themselves; the relationship between indicators and constructs is then viewed as a mereological relation (indicators are part of the network that is the construct) rather than as a causal one (Schmittmann et al., 2013). As will be discussed later, latent variables can be incorporated into the network model, although the network model does not assume latent variables are the explanation for our observations.

What is important to note early on in this chapter is that these models, even though they each imply a specific causal structure and can certainly be used as causal models, are not always used as such. In fact, it is quite common for a psychometrician or a user of psychometric models to not explicitly concern oneself with causality or explaining human behavior, but to focus on prediction or data summarization. The purpose of this chapter is not to prescribe a causal interpretation of these models, or to debunk other interpretations, but to elaborate on the causal structures that these models imply and point out the possibilities and limitations that causal modeling in psychometrics creates. First, we will explain the basic ideas and concepts associated with the reflective, formative, and network models. Second, we will discuss some of the empirical consequences of each of these models, and lastly, we will evaluate a number of problems and solutions that each of these models generates.

4.2 Three Models for the Relation Between Constructs and Measures

In the following three sections, we will discuss three types of psychometric models, each of which implying a different linear causal structure[9]. These models are strongly represented in the psychometric literature and are known according to a wide variety of names. In this chapter, we will refer to the variables that are not observed as 'latent variables', and to the observed variables as 'indicators'. Note that in the literature, latent variables are also referred to as latent traits, unobserved variables, theoretical constructs, and even random effects. The indicators are also often referred to as observed variables or manifest variables.

4.2.1 Reflective Measurement Models

The latent variable modeling tradition originated with the pioneering work of Charles Spearman on human abilities (1904) and was extended by Louis Leon Thurstone's model

[9] Note that in this chapter, we only discuss models that imply a linear causal structure. Nonlinear models, like nonlinear structural equation models, fall beyond the scope of this chapter. We refer to Schumacker & Marcoulides (1998) for a comprehensive overview of nonlinear structural equation modeling, and to Heath (2000) for information on nonlinear methods in psychology.

of multiple factor analysis (1931). This approach has come to be known as *the common factor model*. The common factor model was the first to posit a linear causal relation between a continuous latent variable and a number of continuous indicator variables. In Spearman's case, the latent variable represented general intelligence, and the indicator variables represented test scores on mental ability tests. In the past century, a great many extensions of this basic model have been proposed, which can be understood as variations on the same principle. These extensions mainly differ from the common factor model in terms of the distributional assumptions underlying the model. For instance, if the indicators are categorical rather than continuous, the model translates into a so-called item response theory model (Mellenbergh, 1994b), conceptualizing the latent variable as a categorical variable leads to the latent profile model (Lazarsfeld & Henry, 1968). The discussion below will, for clarity, be cast in terms of the common factor model, but the general issues discussed can be translated to any model that is used to represent a causal relation between constructs and indicators (Borsboom, 2008).

As mentioned earlier, the common factor model is often interpreted as a reflective measurement model, meaning that the variance in the observed scores on the indicators is considered to be caused by variance in the construct (represented in the model as a latent variable). Consider a latent variable η_{1i} and a set of P 'effect indicators' y_{1i} to y_{Pi} for a person i. Furthermore, let λ_{P1} denote the coefficient that gives the expected impact on the indicator y_{1i} as a result of a one-unit change in the latent variable η_{1i} (e.g., in factor analysis, this is the factor loading). In a reflective model, all indicators are a linear function of the latent variable and independent residuals, ε (Bollen & Bauldry, 2011):

$$\begin{aligned} y_{1i} &= \lambda_{11}\eta_{1i} + \varepsilon_{1i} \\ y_{2i} &= \lambda_{21}\eta_{1i} + \varepsilon_{2i} \\ &\vdots \\ y_{Pi} &= \lambda_{P1}\eta_{1i} + \varepsilon_{Pi} \end{aligned} \qquad (1)$$

Figure 1A is a visual representation of a reflective model with four indicators. The common factor model describes how one or more latent variables cause the correlations between a set of indicators. This property follows from the regression equations above, from which we may derive that the correlation between two indicators y_j and y_k is equal to $\lambda_{j1}\lambda_{k1}$, the product of their factor loadings. Under the standard assumption in factor analysis that both the indicators and the factors have variance equal to 1, and that specific factors (the error terms) do not correlate with common factors, it also follows from the equations in (1) that the factor loadings are equal to the correlation between an indicator and the common latent variable.

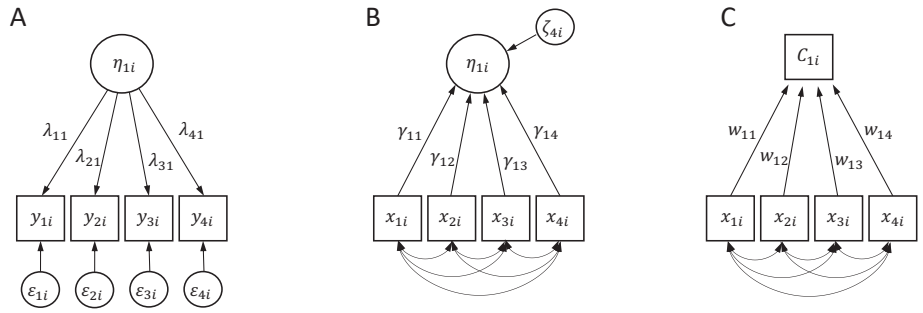


Figure 1. 1A represents a reflective model, 1B represents a formative model, and 1C represents a composite model.

Although not often stated explicitly, when used as a reflective measurement model, the common factor model employs what is known as the *principle of the common cause* (Reichenbach, 1956; Arntzenius, 1993). This important principle of scientific inference is in fact the centerpiece of the method of exploratory factor analysis (Haig, 2005b). In its simplest form, the principle of the common cause asserts that when two variables correlate highly, there must be a third variable that causes this correlation, unless there are good reasons to believe that alternative causal interpretations are more plausible. In science, this simple version of the principle of the common cause is often extended beyond Reichenbach's characterization to include numerous indicators, and more than one latent variable or factor. The best-known form of factor analysis, common factor analysis, is so-called because the factors are common to the production of the correlated indicators. The word 'common' in its name might also be taken to signal its use of the principle of the common cause (Haig, 2005b).

The principle of the common cause is particularly suited to the purpose of common factor analysis in several ways. First, it can be applied in situations where we do not know in advance how likely it is that the correlated effects result from a common cause. Second, the principle can also be used in situations where we are essentially ignorant of the nature of the common cause. Third, the principle is sometimes used as a guide to explanation, in that we appeal to common causes to *explain* their correlated effects. Finally, the principle appropriately restricts the inferences it sanctions to *correlated* effects and *common* causes, which is precisely the focus of common factor analysis.

With regard to the first point, it is important to note that factor analysis provides an answer to the question how strong the effect of the proposed common cause is as opposed to other, incidental effects, among which measurement error. From the equations in (1), it can be deduced that the variance accounted for is equal to the sum of squares of the factor loadings, divided by the number of variables. The third point just made also deserves

further comment, for the common factors in a common factor model are characterized obliquely in terms of their presumed effects under specified conditions. They are not characterized directly by spelling out their processes as causal mechanisms. This can be seen in the case of Structural Equation Modeling (SEM). This form of modeling provides knowledge of causal networks. As such, it does not encourage the development of detailed knowledge of the nature of latent variables. Rather, it specifies the range and order of causal relations into which latent and manifest variables enter. For this type of research, a network theory of causation is needed (Thagard, 1999). Further comments on SEM are offered below.

One of the core assumptions of the common factor model is the *principle of local independence* (Bollen, 2002; Borsboom, Mellenbergh, & van Heerden, 2003; Lazarsfeld, 1959; Lord & Novick, 1968). This principle states that when we condition on the latent variable, the correlations between the observed variables disappear, and these observed variables become statistically independent. In other words: given the latent variable, the observed variables no longer correlate. Cognitive ability tests correlate highly because the scores on these tests are affected by one latent variable. As a result, in applications of the common factor model, it is usually assumed that the indicators have no direct effects on each other, nor do the indicators have an effect on the latent variable (Bollen, 2002). In the following two sections, two models are discussed in which the principle of local independence is violated.

4.2.2 Formative Measurement Models

While it is most common to measure psychological constructs with a reflective model (i.e., the common cause model discussed above), it is also possible to construct a model in which the indicators are causes of the latent variable instead of its effects. Consider a researcher who wants to assess the presence of life stress (psychological stress, as caused by major life events), and constructs a questionnaire that addresses the number of life events a person has experienced in the past year (e.g., changing jobs, getting married or divorced, and moving to a new home; Bollen & Bauldry, 2011). In this case, the relation between the construct (life stress) and its indicators (life events) is not plausibly constructed as one in which the construct causes the observations. In fact, it is more plausible that the latent variable of life stress is caused by its indicators. After all, understanding the relationship between exposure to stress and its indicators as a reflective model would imply that exposure to stress *induces* correlations between indicators such as 'moving to a new home' and 'getting married', which seems highly unlikely. Furthermore, it implies that conditioning on life stress would render the indicators independent. This is also unlikely because if one knows that a person has high life stress (conditioning), then knowing that a person did not move to a new home and has not gotten married or divorced *increases*

the probability that the person changed jobs – after all, the life stress had to come from somewhere. Thus, rather than local independence, one would expect local dependence in this case.

Models in which the latent variable is a common effect of its indicators rather than a common cause are called *formative* models (Edwards & Bagozzi, 2000). Whereas in a reflective model with 'effect indicators', the variance among indicators reflects the latent variable, in the formative model, the variance in the latent variable is produced by variance in the 'causal indicators'. In a formative model, as represented in Figure 1B, the latent variable η_{1i} is a linear function of the causal indicators x_{1i} to x_{Pi} and a disturbance term, ζ_{1i}:

$$\eta_{1i} = \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \ldots + \gamma_{1P}x_{Pi} + \zeta_{1i} \qquad (2)$$

where γ_{1p} denotes the coefficient that gives the expected impact on η_{1i} as a result of a one-unit change in x_{1i}. For example, x_{1i} could represent 'getting married', and x_{2i} could represent 'changing jobs' for a person i, while η_{1i} represents the amount of exposure to stress for this person. In this case, γ_{11} represents the dependence of exposure to stress on getting married, while γ_{12} represents the dependence of exposure to stress on changing jobs. Figure 1B is a visual representation of a causal indicator model with four indicators. Note that in Figure 1B, the indicators are explanatory variables and thus have no residual, whereas in Figure 1A the latent variable is an explanatory variable and therefore has no residual (i.e., disturbance term).

In this example, ζ_{1i} represents either all unidentified causes of exposure to stress or alternatively, is often understood as the variability in exposure to stress that is not explained by the indicators, which could also include random variability. Although these two understandings of the disturbance term are theoretically quite different, they are statistically equivalent and are used arbitrarily. The choice of interpretation of this disturbance term depends on whether one holds a deterministic view, in which each variable is determined by a finite set of causes that can be captured in ζ_{1i}, or whether randomness is inherent to the variable η_{1i}. A debate about the interpretation of ζ, however, is beyond the scope of this chapter.

Strictly speaking, not all formative models are causal indicator models as the distinction between composite models and causal indicator models makes clear. In a composite model, multiple indicators also form a variable, but instead of being a cause of that variable, the indicators *constitute* this variable. This variable is not latent but a linear composite of its indicators (Bollen, 2011; Bollen & Bauldry, 2011). Put differently, the composite variable is completely determined by its indicators and does not have a disturbance term. In this model, the composite C is a function of its indicators x_{1i} to x_{Pi} without any residual:

$$C_{1i} = w_{11}x_{1i} + w_{12}x_{2i} + \ldots + w_{1P}x_{Pi} \qquad (3)$$

Figure 1C is a visual representation of a composite model with four indicators. The parameters w_{11} to w_{1P}, do not represent causal relationships between a latent variable and its indicators, like γ_{11} to γ_{1P}, but denote weights that can be chosen based on some rule. An example of a rule is that the weights are chosen in such a way that the variance in C_1 is maximized. An alternative rule is that weights are selected in such a way that the resulting composite forms the best predictor for some outcome variable.

An example of a composite is the sum score of a set of items on the same scale, in which the sum score is a composite of these items and in which all the items are assigned the same weight: 1. In this case, the latent variable *supervenes* on the indicator scores in the sense that differences in the indicator scores are a necessary but insufficient condition for differences in the latent variable score (Kievit et al., 2011). The relation is not typically seen as causal, because the indicators are not ontologically distinct from the composite so that the relation between the two does not respect the requirement that causes and effects should be separately identifiable (Markus & Borsboom, 2013).

4.2.3 Network Models

Although reflective and formative models are sensible psychometric strategies in many situations, there are also cases where their application is fraught with difficulties. One example of this is the relation between mental disorders and their symptoms in diagnostic systems, such as the DSM-5. Symptoms of depression include insomnia, fatigue, and concentration problems. A reflective model would hold that the correlations between these symptoms are spurious, as they arise from the common influence of a latent variable, but this is unlikely; these symptoms are more likely to feature causal relations among themselves (e.g., insomnia → fatigue → concentration problems). A formative model would suggest that the symptoms are causes of depression, but this is also unlikely because the symptoms are not distinct from depression (one cannot identify depression independent of the symptomatology; Borsboom, 2008; Borsboom & Cramer, 2013). Thus, neither reflective nor formative models sit well with this type of construct, for which causal relations between the indicator variables are both plausible and important.

Recently, it has been proposed that this type of situation requires a different way of thinking about the relations between indicator variables, namely, that these should be seen as variables that are intertwined in a causal network (Cramer et al., 2010; Borsboom & Cramer, 2013). As such, the indicator variables can be subjected to network modeling. An example of a network model consisting of four observed variables (y_{1i} to y_{4i}) is given in Figure 2.

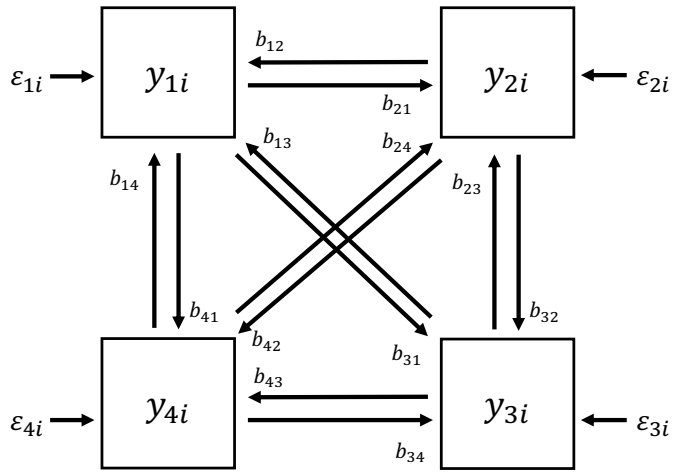


Figure 2. A theoretical network model for four observed variables (or nodes).

A network model encodes the relations between indicator variables in a graph, where nodes represent variables and connections between nodes (edges) represent statistical associations (like correlations or partial correlations). Figure 2 shows a theoretical network model. A network graph based on empirical data will often show edges varying in thickness and sometimes edges will be absent. For a partial correlation network applies that, if a direct connection between any two nodes is missing, this means that the relevant nodes are independent given the other nodes in the system. Statistically, a network that has these properties is called a Markov Random Field (MRF; Kinderman & Snell, 1980; Epskamp, Maris, Waldorp, & Borsboom, 2018).

The idea that the network structure involves direct effects between indicator variables is reflected in the way the model is statistically represented. In particular, one way of fitting the network model to network data is by means of a system of regression equations, where each of the nodes is regressed on all others:

$$
\begin{aligned}
y_{1i} &= b_{12}y_{2i} + b_{13}y_{2i} + \cdots + b_{1P}y_{Pi} + \varepsilon_{1i} \qquad (4)\\
y_{2i} &= b_{21}y_{1i} + b_{23}y_{3i} + \cdots + b_{2P}y_{Pi} + \varepsilon_{2i}\\
&\vdots\\
y_{Pi} &= b_{P1}y_{1i} + b_{P3}y_{3i} + \cdots + b_{P(P-1)}y_{(P-1)i} + \varepsilon_{Pi}
\end{aligned}
$$

In these equations, b_{jk} indicates the regression weight of node k on node j, and ε_{pi} indicates the residual of node p for person i. Unlike the residuals or unique variances in the reflective model, the residuals in equations (4) can in fact be correlated. A network graph in which edges represent the relative importance of a given variable as a predictor of another variable is called a *relative importance graph*. The thickness of the arrow then denotes the strength of the regression weight, which is a quantification of the contribution of a given regressor to the prediction of the criterion (e.g. the lmg metric). A regression weight of zero indicates that predictor variable k has no specific contribution to the prediction of the dependent variable j, given the other variables, and thus the arrow between j and k can be deleted.

It is of some historical interest to mention here that the system of regression equations (4) was proposed and studied already in great detail by Guttman (1953, 1960)[10]. The system provided his alternative approach to factor analysis, called *image analysis*. Since these papers are quite tough reading, we refer for a better digestible summary of image analysis to Mulaik (1972, pp. 186-204). The predicted value of variable j by a weighted combination of the other variables is called the *image* of variable j, and the set of residuals ε_{ji} is called its *anti-image*, which is uncorrelated with its corresponding image. As noted by Mulaik (1972), "Guttman's model of image analysis might have remained a mathematical curiosity, considering the popularity of the common-factor-analysis model, had Guttman not demonstrated over the past 20 years profound relationships between total-image analysis and common-factor analysis." (pp. 191-192).

Aided by a group of prominent psychometricians, Kaiser (1970) codified best practice rules for doing exploratory factor analysis in his Presidential Address to the Psychometric Society, under the heading 'Little Jiffy'. This set of rules still appears to be often used in applications until the present time. The following two Little Jiffy rules are a direct consequence of Guttman's image theory: 1. Use the squared multiple correlation of each regression equation in (4) as the best estimate of the communality of each variable in the factor analysis model (as advised in Guttman, 1956), and 2. Determine if the selection of variables is suitable for a common-factor analysis in the first place by calculating the Measure of Sampling Adequacy (MSA). This measure is based on the sum of squares of the correlations between the residuals ε_{ji} and ε_{ki} across all pairs of variables in (4) (actually, a normalized version called KMO-MSA is used in present-day software, following the update in Kaiser and Rice, 1974). So indirectly, the impact of image analysis on the practice of factor analysis has been rather strong, as it became embedded in Little Jiffy.

[10] The connection between Guttman's work on image analysis and the system of regression equations (4) is based on Heiser (2017).

It is notable that, in contrast to the cases of reflective and formative models, network models do not explicitly represent the construct as a separate term in the model equations. Whereas in the reflective and formative model, the theoretical construct is represented by the symbol in the model equations, in the equations that characterize the network model the construct is represented as a network architecture. Accordingly, although the model is built on the idea that indicator variables are causally related, the relationship between the construct and its indicators is not itself causal in a network model. Rather, this relation is mereological: the indicator variables are *part of* the construct encoded in the network structure (Cramer et al., 2012; Schmittmann et al., 2013). For instance, DSM-5 symptoms like those in Figure 2 are part of the depression construct rather than its cause or effect.

This reflects an important difference between the characterization of constructs using network models versus formative models. The representation of a construct as a network is naturally compatible with the use of composite scores (e.g., the total number of symptoms present) as an indication of the state of the network for a given individual, and in this sense, supports the use of a composite model. However, in the composite model, the construct is a direct function (e.g., a weighted sum) of the variables used to construct it, whereas in the network, the construct is a set of variables together with a set of edges connecting them. A second way in which the network model goes beyond the formative model is that, in the formative model, the set of relations between the constituent indicator variables is ignored; in the network model however, modeling these relations is essential.

4.3 Empirical Consequences of the Different Models

The most obvious difference between the formative and the reflective model is the direction of the causal relation between indicators and the latent variable. But the consequences of this difference in causal direction outreach the interpretation of the latent variable as either a cause or an effect.

The reflective interpretation of the common factor model is that of a common cause model, in which a latent variable is identified as a common cause to the indicators. A common cause model has a number of implications about the covariance structure between the indicators. First, a common cause implies that the indicators share variance. Put differently, because the indicators rely on the same cause they are expected to correlate with each other. Moreover, the principle of local independence implies that the covariance structure among effect indicators is such that a single dimension can be abstracted that has the property that when one conditions on this dimension, the covariance among the effect indicators disappears. These implications result from the rank constraint that a reflective model imposes. That is, the reflective model specifies that the covariance matrix is the sum of a diagonal matrix and a matrix that is constrained to be low-rank.

For example, for a one-factor model, we have already seen that the correlation between two indicators y_j and y_k is approximated by the product of the factor loadings: $\lambda_{j1}\lambda_{k1}$. More generally, for a factor model with q factors, the empirical correlations are approximated by the sum of q products of factor loadings: $\lambda_{j1}\lambda_{k1}+\lambda_{j2}\lambda_{k2}+ \ldots +\lambda_{jq}\lambda_{kq}$. Generally, for a factor model with q factors, the model implied covariance matrix is the sum of a diagonal matrix and a matrix that has rank q (Shalizi, 2013). Because the one factor model implies this rank constraint, it is not possible to write any arbitrary covariance matrix in terms of a one factor model. And because some, but not all, covariance matrices comply with a one factor model, the one factor model has testable implications for the data (Shalizi, 2013).

Factor analysis is thus the act of approximating a sample covariance matrix with a low-rank matrix[11], which can be done with maximum likelihood estimation. An alternative way to test the rank constraint of a reflective model is based on the fact that a matrix of rank one implies that some so-called tetrads equal zero. A tetrad refers to the difference between the products of two pairs of covariances among four random variables (Bollen & Ting, 1993). How to compute such tetrads is beyond the scope of this chapter, but we refer the reader to Bollen and Ting (1993; 2000) for an overview. The tetrads that equal zero are called vanishing tetrads and factor models with different ranks, and different numbers of variables imply different vanishing tetrads that can be tested.

The reflective model thus imposes constraints on the covariance structure and as such, offers a way to test the likelihood of the reflective model, given the observed covariance structure. The formative model itself, as presented in Figure 1B, does not impose such constraints on the covariance structure among the indicators. The relations between causal indicators can be negative, positive, or anything in between. Because of the lack of constraints, the model in 1B is not identified. It is only when some effects of the latent variable are introduced in the model, that the model again starts to impose testable constraints on the covariance matrix. The model in Figure 3 has a set of causal indicators as well as some effect indicators and is called a Multiple Indicators MultIple Causes (MIMIC) model (Jöreskog & Goldberger, 1975).

The covariance matrix of a MIMIC model includes both causal indicators and effect indicators. The effect indicators are necessary to identify the model as the model implies testable constraints on the data. The MIMIC model implies that the effect indicators are independent given the latent variable, and that for any combination of a causal indicator and an effect indicator, these variables are independent, given the latent variable, as the latent variable mediates these relations (see Figure 3). Additionally, the MIMIC model implies additional constraints on the covariance matrix of observables, because all relations

[11] Strictly speaking, the sample covariance matrix is approximated with a low-rank matrix with a correction, namely the diagonal matrix of unicities.

between the causal indicators and effect indicators go via the same latent variable, and correlations between observables should be proportional to the size of the causal effects that are propagated through the latent variable. These constraints make it possible to fit the model to the data. Another way to evaluate whether the MIMIC model is correctly specified is that the model implies that the coefficients for the influence of the causal indicators on the latent variable are stable for different effect indicators. That is, adding effect indicators to the model, or replacing effect indicators with other effect indicators of the same factor, should not influence the estimated γ coefficients of the causal indicators. After all, if the latent variable is properly identified, the coefficients between the same set of causal indicators and the latent variable should not change as an effect of what outcome measure (or effect indicator) is used. When the estimated γ coefficients change with different effect indicators in the model, this means that the latent variable is unstable and thus that the model as a whole suffers from misspecification.

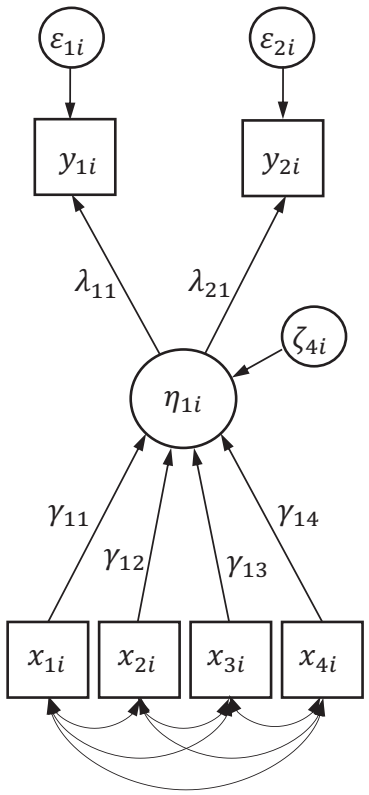


Figure 3. A Multiple Indicators MultIple Causes (MIMIC) model with four causal indicators and two effect indicators.

So far, networks have mainly been used in an exploratory manner. This means that instead of hypothesizing a specific network structure and testing the conditional independencies that are implied by this network on the data, the data is used to identify conditional independencies and construct a network accordingly. However, given a specific network model, it is also possible to derive testable implications for the data. For example, if a specific network is hypothesized in which some nodes are not connected with an edge, then this network model implies that the variables that correspond to these missing edges are conditionally independent, given the other variables in the network, which is in itself a testable hypothesis. If the hypothesized network consists of two clusters that are connected via one single node, this means that any set of two nodes of which one is part of the one cluster and one is part of the other cluster are conditionally independent given the node that connects the two clusters. Recently, tools have been developed that can test such hypotheses, in a similar way that confirmatory factor analysis assesses model fit (e.g. psychonetrics, Epskamp, 2020).

4.4 Controversies

So far, we have described three types of models and given an overview of the core essentials pertaining to each of them. However, neither latent variable models nor network models are unproblematic. In this section, we will discuss four controversies surrounding these models and suggest further reading.

4.4.1 The Causal Status of Individual Differences

Latent variable models are typically fitted to datasets that contain different variables measured in a large number of people for a single time point. Because of the fact that no temporal information is present in such data, the only source of variation that the model can utilize is variation between individuals. As a direct result of this issue, both the latent and observed variables exclusively refer to dimensions of individual differences. The question whether such dimensions can, in fact, be viewed as causally effective entities has generated a fair amount of interest (Molenaar, Huizenga, & Nesselroade, 2003; Borsboom et al. 2003; Hamaker, 2012; Hamaker, Kuiper, & Grasman, 2015; Borsboom, 2005, 2008, 2015; Markus & Borsboom, 2013).

By assumption, the reflective measurement model introduces individual differences variables as causal factors. The controversy at hand focuses on the question of whether, and if so, how this interpretation pans out at the level of the individual person. In particular, it is not clear whether, *if* one espouses the causal interpretation of that individual differences variable itself, one is *also* committed to a causal interpretation at the level of the individual person. If such a necessary connection existed, that would seriously complicate the applicability of the reflective measurement model to standard psychometric practice.

For example, even though it seems sensible to say that individual differences in the *g*-factor cause individual differences in IQ-scores (Jensen, 1999), it seems considerably less sensible to say that a person's position on the *g*-factor causes his response to the IQ-item "who wrote the *Ilias*?" (Borsboom et al., 2003). There are two reasons for this.

The first issue is epistemological in nature. It can be shown statistically that intra-individual causal relations need not be isomorphic to the regression equations that make up psychometric models. That is, it is possible to generate data in which none of the individuals in a population are described by the psychometric model for the individual differences variables (Molenaar, 2004; Hamaker, 2012; Hamaker, Kuiper, & Grasman, 2015). The psychometric model might only describe what happens at the aggregate level, so when averaged across individuals, which might not correspond to what happens at the individual level. This issue has also been raised in the context of network analysis (Bos & De Jonge, 2014; Bos & Wanders, 2016; Van Borkulo et al., 2015; Van Borkulo, Borsboom, & Schoevers, 2016).

This compromises the evidence for causal interpretations of individual differences variables. In the case of *g*-factor, for instance, one could hold that there is some reasonable evidence for the hypothesis that there is indeed a single dimension of individual differences that permeates scores on different cognitive tests (Jensen, 1999), but there is almost no evidence for that *g*-factor being in some sense causally operational within the individual person – simply because all of the evidence for the *g*-factor is based on the analysis of individual differences (Borsboom & Dolan, 2006). We would have to conclude that (a) there is no direct evidence for this within-person hypothesis at all, and (b) we have not the slightest idea of even how to gather such evidence, because the theory of the *g*-factor is silent on whatever happens at the level of the individual. It seems, therefore, that the evaluation of variables like the *g*-factor as bearing causal implications at the level of individual persons is not only unrealistically strong but also beside the point.

Second, there is a problem of ontology (i.e., concerning the existential status of individual differences). This problem arises because one's position on an individual differences variable would appear to depend ontologically on the actual presence of individual differences, and thus on the presence of other individuals, who together 'generate' or 'constitute' the dimension in question (Borsboom, 2008). However, the concrete behavior of answering an IQ-item surely does not depend ontologically on the presence of other individuals. Therefore, if everyone except you now suddenly died, leaving you the only sentient being in the universe, it seems there would be no individual differences anymore, and thus no *g*-factor for you to have a position on (Borsboom, 2015). But in this case, you would still be able to solve an IQ-item. It, therefore, appears that to invoke individual differences in the *g*-factor as a cause of an individual solving an IQ-item is superfluous. Unless one is willing to make clearly problematic ontological claims (e.g.,

that the dimension of individual differences, or a person's position on it, would still exist even if there were no other people; or that behavior is overdetermined by both individual-level psychological processes *and* individual differences variables) it seems that individual differences variables must be considered causally impotent with respect to individual item responses.[12]

Weinberger (2015) has argued (within the causal interventionist framework of Woodward, 2003) that this situation is inadmissible, and that if a variable such as the *g*-factor is to be causal at all, it must *a fortiori* be causally effective within the individual. The reason is that, in an interventionist account, a causal relation means that to change a person's position on a causal variable would instigate a change on an effect variable, and this cannot be the case in the population if it is not the case for at least a subset of individuals in the population. Borsboom (2015), however, argued that this argument does not work because the interventionist account does, in fact, allow for interventions that change the individual's position on a variable without changing anything about the individual (e.g., by changing the population composition which generates the individual differences).

All in all, this discussion is unresolved, and it is currently unclear whether individual differences variables are plausibly interpreted as causally effective entities, and in what framework such an interpretation is most plausibly cast. Naturally, this discussion carries over to any causal framework involving individual differences, and thus affects both formative models and network models (insofar as these rely on individual differences dimensions), although it has not led to equally vigorous debates in these areas.

4.4.2 Generalization

The tradition of latent variable modeling began as a substantive approach to modeling rather than as a data analytic approach. The common factor model was the direct result of Spearman's construction of a theory of general intelligence. In this approach, latent variable models are representations of causal relations between unobserved psychological factors and item responses. Specifically, latent variables are assumed to *explain* these responses. For this reason, latent variable models can play an important role in building scientific theories. One of the purposes of science is to find justified explanations for our observations, and in many fields, latent variable models have done just that. As we have

[12] It is also possible to understand individual differences as a personal trait that does not depend on other individuals. In the case of educational measurement, the individual's scores are compared to an objective age-appropriate norm which is determined by the kind of tasks the individual has to be able to pass, not to the level of other people's traits. When individual differences are understood as personal traits, the problem of ontology can be dropped.

seen earlier, latent variable models have contributed to theory construction in the domains of intelligence, personality, and psychopathology.

One advantage that latent variable models have over models that are based on only the indicators is that they can be used for building scientific theories about the population (Bentler, 1980). There is a wide range of tests available to measure all sorts of constructs, and which test is used in a specific study depends on the preference of the researcher, or simply on the availability of the tests. So when a model is constructed based on relations between only observed indicators, without assumed relations between these indicators and latent variables, it is likely that this model is strongly biased due to preferences for certain measures and not others. In latent variable models, however, it is assumed that it does not matter what kind of tests are used to measure the latent variable: the latent variable is assumed to exist independently of the indicators. A second reason why building population theories based on manifest variable models is risky is that these models do not make a distinction between error and meaningful effect (something latent variable models explicitly do). Whereas in latent variable models, it is assumed that parameter estimates are invariant over different populations, this can hardly be the case with models that are only made up of manifest variables unless the measurement error is exactly the same for different populations. Therefore, latent variable models form a useful tool to build general scientific theories pertaining to the population as a whole.

In the common factor model, in which a single common factor explains the covariation between the observed variables, only the causal relations between one latent variable and a set of observed variables are modeled. This model is often used as a measurement model, in which one latent variable is measured by a set of items or tests. Throughout the years, it has become possible to also model the causal relations between multiple common factors and the observed variables (Thurstone, 1931), and later on, also the causal relations between latent variables themselves (Jöreskog, 1970). The latter is also known as Structural Equation Modeling, or SEM. Subsequently, many psychological models consisted of a *measurement* part and of a *structural* part. The measurement part represents the relation between latent variables and a set of test scores or item scores, and the structural part represents the relations among the latent variables themselves. SEM models thus represent hypothesized causal relations between a set of latent variables, each causing variance in a set of indicators.

Similar to some factor models (a very early example being Thomson's bonds model of intelligence; Thomson, 1916), SEM models are often hierarchical, meaning that the model incorporates multiple-order factors. The measurement model is thus nested in a higher-order model. Variation in the observed variables is explained by a latent variable, which is, in turn, explained by a higher-order variable. To evaluate how well the data match the hypothesized structure, the model's *goodness of fit* can be assessed. Through

cross-validation or replication of the model, one gathers support for the fit of the model to the relevant empirical evidence. In this way, SEM has become a useful tool to test causal hypotheses among latent variables. All together, factor analysis, and its extensions, such as SEM, enable researchers to model different types of causal relations among observed and latent variables (Bollen & Pearl, 2013). They can therefore serve as an important method for theory construction.

4.4.3 Interpretational Confounding

Interpretational confounding is the phenomenon that the meaning of a latent variable changes as a result of a change in outcome variables that are used to identify the full model. This can be best explained by taking a look at Figure 4. Figure 4a represents a model with three latent variables η_1, η_2 and η_3, that are related in such a way that η_1 is hypothesized to cause η_2 and η_3. In this model, the relations between the latent variables are structural relations whereas the relations between the latent variables and their indicators are measurement relations. The indicators y_1 to y_4 identify η_1, the indicators y_5 to y_8 identify η_2 and the indicators y_9 to y_{12} identify η_3.

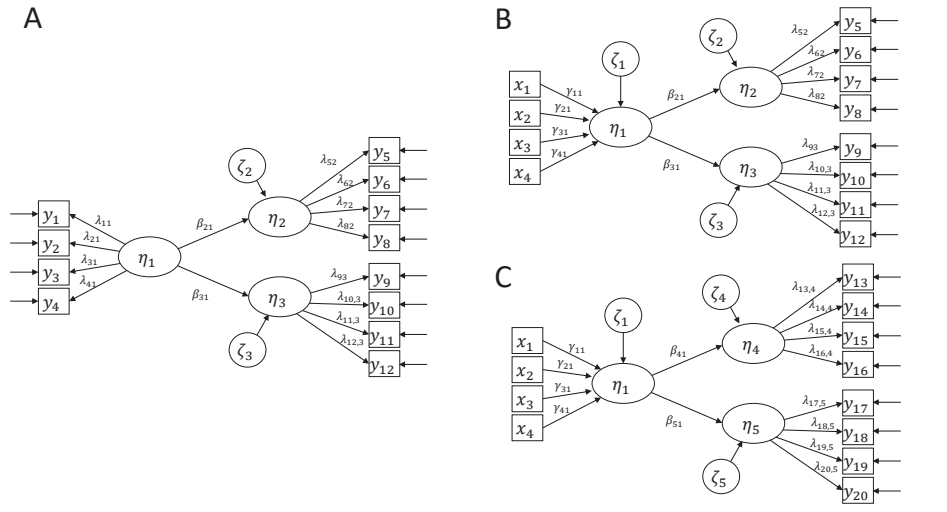


Figure 4. In 4a, all three latent variables η_1, η_2 and η_3 are measured with a reflective model, and thus each latent variable is identified by its effect indicators. In figure 4b and 4c, η_1 is measured with a formative model and η_1 is therefore identified by the causal indicators and the two outcome variables (i.e., η_2 and η_3 in 4b and η_4 and η_5 in 4c).

Suppose that one of the latent variables, η_1, forms a causal indicator model rather than a reflective model, such as in Figure 4b. In this model the relations β_{21} and β_{31} are no longer pure structural relations, but also identify η_1 via η_2 and η_3. As a result, in case η_2 and

η_3 are replaced by two different factors η_4 and η_5 (Figure 4c), the meaning of η_1 changes, because η_1 and η_2 will identify a different factor than η_3 and η_4 (i.e., η_1 in Figure 4b is not η_1 in Figure 4c). Because η_1 in Figure 4b is different from η_1 in Figure 4c, the coefficients γ_1 to γ_4 of the causal indicators will also be different for the model in Figure 4b and the model in Figure 4c. As a result, researchers may believe that they measure the same latent variable η_1 because they use the same set of causal indicators to measure η_1 while, in fact, the meaning of this variable changes as a result of what outcome variables they use to identify the full model. For this reason, Howell, Breivik, and Wilcox (2007) conclude that the formative model is a less attractive measurement model than the reflective model as it is more prone to interpretational confounding. As mentioned earlier, there is a debate going on about whether formative models are measurement models, and the fact that causal indicators alone cannot identify a latent variable is an important argument in this debate. We refer the reader to Howell et al. (2007) and Bainter & Bollen (2014) for a deeper understanding of both sides in this debate.

In composite models, effect indicators are not needed to identify the composite because the composite is nothing more than the linear combination of the composite indicators. However, when composites are used for prediction, the outcome variable that is predicted by the composite is often used to estimate the weights of the composite indicators, such that the resulting composite optimally predicts this outcome variable. For example, Socio-Economic Status (SES) can be used to predict experienced happiness, and the weights of the indicators of SES can be estimated such that SES explains the most variance of experienced happiness. However, when another researcher decides to use SES to predict another variable than experienced happiness (e.g., self-confidence), this researcher will estimate different weight parameters even though this researcher might use the same set of indicators. Now, since SES is fully determined by these weights and the indicators, SES is something different when used to predict experienced happiness than when SES is used to predict self-confidence. This is another example of interpretational confounding, because although the construct SES will probably be interpreted in the same way across these two different studies, the meaning of SES actually changed as a result of which outcome variable is used.

In contrast, if η_1 in Figure 4c were to be measured with a reflective model like in Figure 4a, the meaning of η_1 would not differ over Figure 4b and 4c. The interpretation of η_1 thus does not depend on the outcome variables η_2 and η_3. As mentioned in the previous section, a property of the reflective model is that it assumes that the latent variable can be identified independently of which set of indicators is used, as long as the indicators are an effect of the latent variable. However, this is only the case if all indicators are indeed all effects of the same single variable. When this assumption is violated, interpretational confounding can take place in the reflective model as well.

For example, consider a set of six indicators of which it is believed that they are caused by the same latent variable while actually they are caused by two different variables that correlate highly or have some overlap in meaning. Now suppose that the indicators γ_1 to γ_3 are an effect of the first latent variable η_1 and the indicators γ_4 to γ_6 are effects of η_2. The shared variance of these six indicators reflects the correlation between the two latent variables (assuming the factor loadings are roughly equal). However, obviously, by taking out indicators of η_1, the meaning of the factor drifts to η_2, while taking out indicators of η_2 drifts the meaning of the common factor to η_1. As a result, since the researcher believes that the shared variance of γ_1 to γ_6 reflects one latent variable, it is assumed that taking out indicators or adding indicators does not change the meaning of the latent variable, even though it does change the meaning of the latent variable in this example. Therefore, in order to prevent interpretational confounding when taking out or replacing indicators, it is important to evaluate the assumption that all effect indicators in the reflective model are effects of one single latent variable. Also, in this case, observing that the factor loadings change as a result of adding or removing indicators signals that the meaning of the latent variable changes.

4.4.4 The Relationship between Causal Models and Data Models

In science, models are often used for representing causal relations and improving one's understanding of them. 'Scientific models' is an umbrella term that covers all kinds of models (Frigg & Hartmann, 2005). An important distinction is that between a substantive model and a data model. A substantive model incorporates parameters that represent a certain real-world mechanism. For the construction of these models, scientific theory is used to impose meaning on the individual model parameters (Haslbeck, Ryan, Robinaugh, Waldorp & Borsboom, 2020). Because substantive models are often used for representing mechanisms, substantive models are often *causal models*. More specifically, causal models are formal theories that state the relationships between precisely defined variables (Blalock, 1985). Data models, on the other hand, only model the statistical dependencies in the data and are not direct representations of substantive theory (Moneta & Russo, 2014; Van Fraassen, 2008). One example of a data model is a regression line through a set of data points. This regression line is a way to represent the statistical dependencies in the data about how x and y are related, but it does not bestow any substantive, or causal, meaning on x and y. In this model, it is clear that x and y are correlated, but the substantive interpretation of their relationship is not specified. Applications of data models are, for instance, models that are merely used for prediction purposes, which do not intend to appeal to actual causal mechanisms.

Whether models can actually refer to real-world mechanisms at all has been a topic of debate. According to some, models are indeed capable of approaching the truth,

whereas others find this perspective too liberal. In their view, models are simply tools that are useful for prediction, but they do not have any claim on reality (Fine, 1993; Barberousse & Ludwig, 2009). Whether latent variable models are causal models has both been advocated and disputed over the last decades (Block, 1974; Maraun, 1996; Mulaik, 1996). Regardless of this dispute, latent variable models are in practice often *treated* as causal models. And as the section on empirical consequences has described, there are reasons for assuming one causal structure over the other. However, in practice, it is possible to use latent variable models for something other than causal modeling and treat the model as a data model. A famous application of the common factor model is data reduction, in which the user is interested in reducing a large number of variables to a smaller number (common factors). It is, therefore, a very strong position to hold that the common factor model and other latent variable models are inherently causal of nature. This would imply that the model can only be used when a certain causal relation is expected. In fact, nothing stops anyone from using the model for other purposes as well. It is therefore useful to distinguish the causal reading of a model from the model itself. The model itself does not imply a causal reading on its own; the interpretation is *added* to the model rather than implied by the model. The common factor model can be used to represent a mechanism that is best described by a common cause structure, but may also be used as purely a data model, used to describe certain statistical dependencies in the data. In a similar vein, the network model can be used to analyze causal relationships among a set of variables, but also for detecting clusters of variables or for visualization purposes. In other words, the intentions of the user are fundamentally important for how the model is used.

4.4.5 A Realist Philosophy of Psychometrics vs. Psychometric Practice

Hood (2013) and Borsboom et al. (2013) state that the reflective models that are so popular in psychometrics are in accordance with a realist philosophy, and Chapters 4 and 5 in this dissertation are largely written in the same tradition. The fact that psychometricians often choose a reflective model over other types of models and have psychological or educational measurement as a goal, uncover realist intentions. Even though there are probably realists among the psychometricians and a realist philosophy is indeed in accordance with psychometric models, an attitude that is agnostic or even weary towards the causal status of models and the existence of latent variables is probably more common among the psychometricians themselves. It is not uncommon for psychometricians to incorporate latent variables in their research, but to consider them merely random effects or dimensions of the data. The realist philosophical approach that is implied by psychometric models thus deviates from what psychometricians are actually doing and how they interpret their results.

The philosopher of science would possibly point out the incongruence between the choice of models and the actual interpretation and use of psychometric models. Why use reflective models at all if prediction or data summarization is the purpose of the study? Why not choose models that do not imply that causal relationship? Though this is an interesting question (which will be elaborated on in the following chapter), an equally interesting question is why there exists such a deviance between the realist philosophy that is implied by the models and the beliefs and practice of psychometricians? Why are psychometricians reluctant to speak about causality in the first place?

Though there is no conclusive answer to this question, we can hypothesize about reasons why this is the case. One possible reason is that history has proven that the unwarranted reification of psychological attributes is not without risk (Danziger, 1997; Evans & Waites, 1981; Gould, 1981). The strong belief in the existence of general intelligence by early psychometricians facilitated scientific racism and had highly unpleasant consequences for people other than wealthy white men (see Chapter 6). With that in mind, psychometricians might prefer to refrain from making too strong statements about real-world entities or causal relationships and rather stick to the technicalities of the models. Considering these models data models rather than causal models is simply the safer choice. A second reason could be that psychometricians have more faith or interest in the data analytical power of models than in the truth-component of psychological theory. Psychological theory is still relatively underdeveloped and has not yet reached a reliable state. Moreover, there is still no consensus about the ontology of psychological attributes (Goertzen, 2008; Yanchar & Hill, 2003). It is imaginable that in order for psychometric models to be treated as causal models, psychometricians would require stronger evidence for the actual existence of psychological attributes.

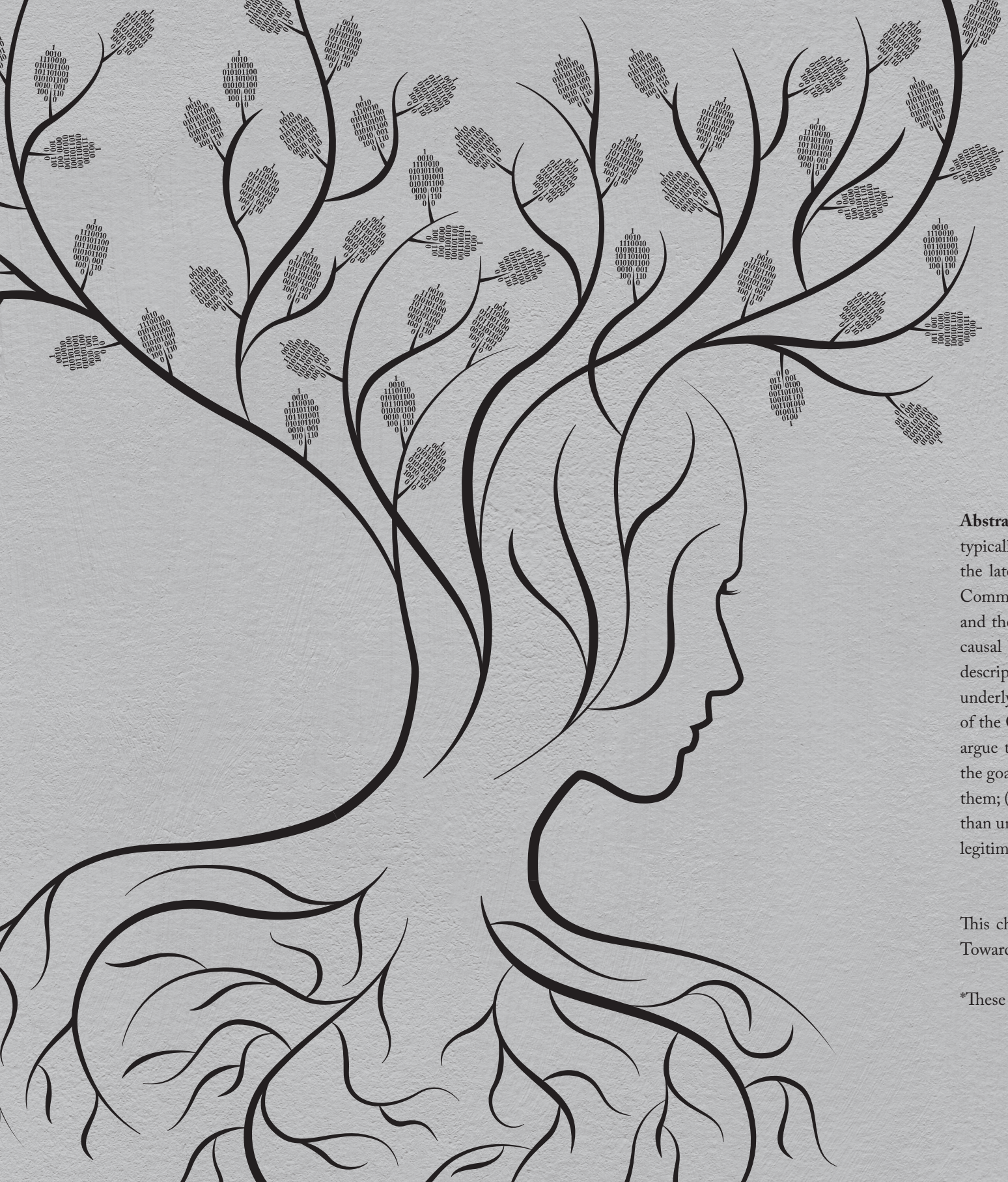
So even though this chapter and the next are written with a realist framework in mind, it is important to take note of the fact that psychometricians do not completely abide what a realist philosophy prescribes, and possibly have valid reasons for this. A study into why a realist philosophy of psychometrics and psychometric practice are at odds with one another and how this has come to be would be valuable to increase our understanding of psychometric research and is a topic worth investigating.

4.5 Conclusion

In this chapter, we have provided an overview of a number of important methods for modeling causal relations in psychological science. In psychology, latent variable models have dominated causal modeling in psychology for the last century. In the reflective model, a causal interpretation of the common factor model, the latent variable is modeled as the explanation for correlations between the indicator variables. In this model, latent variables are the causes that produce the correlations. In formative models, these

relations are reversed, as now the indicator variables together cause or construct the latent variable. In the formative model, a distinction is made between causal indicator models and composite models. In the causal indicator model, the indicators are causes of the latent variable, whereas in composite models, the indicators constitute the latent variable. Lastly, we discussed network models, in which the causal relations are drawn between the indicator variables only, and latent variables are left out of the equation.

The section on controversies showed that all of these models come with certain problems, and it was in no way our intention to state that using any of these models leads to sound conclusions about causal relations in psychology. On the contrary, the problem of how to establish causal relations was not resolved with the development of either of these models. For instance, it is possible that data support both a reflective and a network model, in which case it is difficult to determine which one is in fact the true model. For example, the correlational structure that is implied by the common factor model with a single factor is also implied by the mutualism model, in which cognitive process mutually interact as opposed to being the effect of a common factor (Van der Maas et al., 2006). As mentioned earlier, some models have certain empirical consequences that can be verified, but preference for either of these models is often based on theoretical considerations. The question whether depression can be considered a latent cause or is constituted by relations between its symptoms, is first and foremost a theoretical question. Latent variable models and the network model provide methods for thinking about different types of causal relations, as they lead to fundamentally different conclusions about the nature of psychological entities and their explanatory role. So even though the models mentioned in this chapter do not conclusively solve the problem how to draw conclusions about causality in psychology, they may contribute to a better understanding of the phenomena in question.



Chapter 5

A Causal Interpretation of the Common Factor Model

Abstract Psychological constructs such as personality dimensions or cognitive traits are typically unobserved and are therefore measured by observing so-called indicators of the latent construct (e.g., responses to questionnaire items or observed behavior). The Common Factor Model (CFM) models the relations between the observed indicators and the latent variable. In this article we argue in favour of interpreting the CFM as a causal model rather than merely a statistical model, in which common factors are only descriptions of the indicators. When there is sufficient reason to hypothesize that the underlying causal structure of the data is a common cause structure, a causal interpretation of the CFM has several benefits over a merely statistical interpretation of the model. We argue that (1) a causal interpretation conforms with most research questions in which the goal is to *explain* the correlations between indicators rather than merely summarizing them; (2) a causal interpretation of the factor model legitimizes the focus on *shared*, rather than unique variance of the indicators; and (3) a causal interpretation of the factor model legitimizes the assumption of local independence.

This chapter is adapted from Van Bork*, R., Wijsen*, L. D., & Rhemtulla, M. (2017). Toward a causal interpretation of the Common Factor Model. *Disputatio, 9*, 581 – 601.

*These authors have contributed equally.

5.1 Introduction

One of the many pursuits of psychology is to establish causal relations between properties of the mind and human behavior. Psychologists are interested in the motivations, cognitive abilities and personality traits that explain why people behave in a certain way. Chapter 4 elaborates on a number of possible models that can facilitate causal modeling in psychology. However, causal modeling in psychology has proved to be anything but easy. Part of the problem comes down to the very nature of psychological constructs. In psychology, the attributes that are used to explain human behavior (e.g., personality characteristics that explain individual differences in behavior) are typically *latent variables*, that is, theoretical constructs that are unobserved (Bollen, 2002; Borsboom, Mellenbergh, & van Heerden, 2003; Hood, 2008). Because latent variables are by definition unobserved, psychological tests are constructed to measure these variables. The observed responses on such tests are believed to reflect the latent variable that underlies them. For example, *intelligence* cannot be directly observed. Yet, it is assumed that intelligence can be measured by administering a set of IQ items to which the responses can be observed. The measurement model on which this mechanism is based is the *reflective model* (Figure 1). The IQ test is supposed to measure intelligence because it is built on the assumption that the responses to the items are a direct *effect* of one unobserved entity, intelligence. Consequently, the variance shared among the test scores is assumed to reflect this theoretical entity. The variance that is unique to each item is assumed to reflect measurement error as well as unique causes of the responses to that item (represented by to in Figure 1). The reflective model is an example of a 'common cause' model because the latent variable functions as a cause of all of the item responses. In the remainder of the paper, we refer to the observed variables (such as the test scores in this example) as the *indicators* of the latent variable.

Latent variable modeling originated with the construction of the *common factor model* (CFM; Spearman, 1904). Spearman observed what is now known as the positive manifold: item responses on a variety of mental ability tests were all positively correlated with one another. Using the principle of the common cause in which a correlation between two variables is explained by a third variable, Spearman abstracted the general factor of intelligence (*g*-factor) from these test scores, arguing that one underlying entity explains the shared variance in all branches of intellectual activity: namely, general intelligence. In the CFM, all indicators are a linear function of the common factor and the indicators are statistically independent conditional on the latent variable. After all, if a common cause explains the covariation between two indicators, the indicators no longer correlate when conditioning on this common cause. So, a CFM of general intelligence implies that all branches of intellectual activity are rendered independent, conditional on general intelligence. This principle is fundamental to Spearman's CFM and to reflective models in general, and is called the *principle of local independence*. The process of fitting a CFM

to data is called *factor analysis*. Factor analysis enabled psychologists to discover and test theories about plausible explanations for human behavior. It is therefore not surprising that CFMs have become very popular, not only in intelligence research, but also in the fields of psychopathology (e.g., Asmundson et al., 2000; Caspi et al., 2014) and personality (e.g. McCrae & Costa, 1987; Musek, 2007).

We reserve the term "reflective model" for the theoretical causal interpretation of the CFM. In contrast to the reflective model, the CFM is not defined as a causal model but is typically defined as the set of equations that equate each indicator X_i to a function of the latent variable η and a unique component ε_i that is typically called *the residual* in factor analysis. For example, consider four indicators X_1 to X_4 of the same latent variable η. Each indicator is a linear function of the same latent variable and a unique residual component:

$$\begin{aligned} X_1 &= \lambda_1\eta + \varepsilon_1, \\ X_2 &= \lambda_2\eta + \varepsilon_2, \\ X_3 &= \lambda_3\eta + \varepsilon_3, \\ X_4 &= \lambda_4\eta + \varepsilon_4. \end{aligned} \qquad (1)$$

These equations are typically graphically represented in the same way as the reflective model (see Figure 1). The covariance matrix of the indicators is a function of the vector of factor loadings, the variance of the latent variable and the covariance matrix of the residuals:

$$\Sigma = \lambda\psi\lambda + \Theta. \qquad (2)$$

Although the equations above are agnostic with respect to causality, we argue that when factor analysis is used to measure a psychological construct, the CFM benefits from being interpreted as a reflective model, that is, interpreting the latent variable as the common cause of the indicators. The reflective model in Figure 1 can be seen as a specific case of the CFM: it includes the CFM equations above, and adds to them a causal interpretation (Bollen & Lennox, 1991; Edwards & Bagozzi, 2000).

Even though latent variable modeling has had a profound influence on psychological research, methodologists and modelers have not reached a general consensus about how to understand the nature of latent variables (Bollen, 2002; Borsboom et al., 2003; Jonas & Markon, 2016). As latent variables are not directly observed, it is typically impossible to perform a manipulation on them to determine (a) whether they exist at all, and (b) whether they really *cause* the observed responses that they are supposed to cause. For

example, it is impossible for us to manipulate a person's intelligence to test whether the answers on an IQ test change as a result of this intervention[13].

It is not only the unobservable character of psychological constructs that makes causal modeling in psychology a complicated endeavour. Factor analysis is most frequently performed on cross-sectional data, gathered at a single time point. Such data cannot be used to distinguish between models that have different causal structures but are statistically equivalent. Whenever a CFM is fit to a dataset, there are alternative causal structures that may equally well have generated the data (Van der Maas et al., 2006). Because of this ambiguity, many psychometricians advocate sticking to a strict statistical interpretation of the CFM, to avoid inferring causality without direct evidence for it. They would argue for a descriptivist approach in which the latent variable merely represents the shared variance among a set of indicators. The descriptivist approach understands latent variables as a parsimonious summary of the data, rather than an underlying cause of the indicators (Jonas & Markon, 2016).

Thus, although the CFM was developed as a model in which the common factor is hypothesized to represent an existing causal entity (i.e., general intelligence) that explains patterns in different branches of intellectual activity (Spearman, 1904), the descriptivist approach views the common factor that is obtained with factor analysis merely as "just a convenient way of summarizing patterns of observed relationships" (p. 91, Jonas & Markon, 2016). Jonas and Markon (2016) argue:

> Reflective latent variable models are agnostic with regard to the nature of the etiological process: this is the heart of the descriptivist paradigm. Reflective latent variable models can describe data generated from any number of etiologic processes; consequently, the form of the latent variable model cannot arbitrate questions of causality. (p. 91-92)

Factor analysis does not test any causal relations but rather absorbs shared variance in a common factor. For any arbitrary set of indicators that share variance, the shared variance will constitute a common factor, no matter what caused this shared variance. However, in order to measure a psychological construct of interest, one does not consider any randomly chosen set of variables, but rather a specific set of variables that are hypothesized to be affected by the construct. The hypothesis that the indicators reflect the psychological

[13] Note that it is possible to hypothesize an intervention on intelligence, e.g., drinking a lot of alcohol (Borsboom, Mellenbergh & van Heerden, 2003). However, since intelligence is only visible via its reflection in the indicators, it is problematic to test whether this intervention actually affects intelligence or whether it instead affects the indicators directly.

construct of interest enables the researcher to interpret the common factor as this construct, by the logic that, *if* the indicators share a common cause, the shared variance among these variables reflects this cause (Edwards & Bagozzi, 2000).

In this paper, we argue against the view that the common factor is, at all times, merely a convenient summary of the data. Instead, we argue for a causal interpretation of the CFM when used to measure constructs, in other words, when the purpose is explanation. When a causal interpretation is not justified, for example because it is unlikely that the construct of interest causes the indicators, other statistical models might be equally or more appropriate (see for example Bollen, 2011; Diamantopoulos, Riefler, & Roth, 2008; Edwards & Bagozzi, 2000). Although a causal interpretation of the CFM is neither always justified nor necessary, it offers several benefits over a purely statistical reading of the model in cases where it is justified. To make this argument we first distinguish between statistical models and causal models, and explain why researchers might be reluctant to interpret the CFM causally. Subsequently, we outline the descriptivist approach in more detail. Finally, we present three arguments that highlight the benefits of a causal interpretation over a purely statistical interpretation of the factor model.

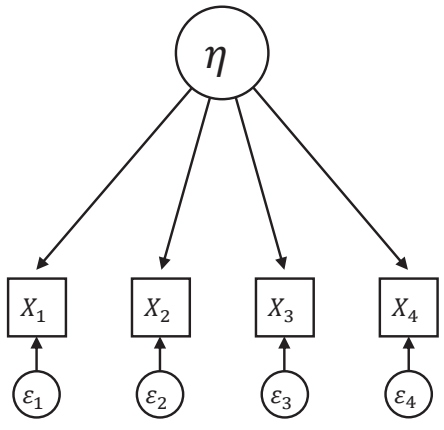


Figure 1. The reflective model. Squares represent indicators (e.g., X_1 to X_4 could be the responses on four different IQ items) and ε represents the latent variable (e.g., intelligence). ε_1 to ε_4 refer to measurement error and unique causes of the indicators. Eliminating these unique factors would imply that the indicators are fully determined by the latent variable, which is typically not assumed for causal processes in psychology.

5.2 Statistical vs. Causal Models

We distinguish between statistical models on the one hand, and statistical models with a causal interpretation on the other. For the sake of brevity, we use 'causal model' to refer

to the latter, though we acknowledge that causal models need not be statistical. A statistical model is a set of probability distributions on some sample space (McCullagh, 2002). This means that no conceptual interpretation is yet included; a statistical model only includes the statistical dependencies among indicators (Moneta & Russo, 2014). In contrast, a causal model has an additional causal interpretation because one or more of the parameters in the model reflect causal relations. A causal model could be seen as a representation of a real-world phenomenon, rather than a model that only plots the relations in the data. We are aware that there exists a large literature on the definition of a model and the distinction between different types of models in philosophy of science (e.g., Van Fraassen, 2008; Frigg & Hartmann, 2005), but a discussion of this literature is beyond the scope of this paper.

Two models that represent different causal relations can be statistically equivalent. For example, many structural equation models are statistically equivalent to a model in which one of their structural relations is reversed (MacCallum, Wegener, Uchino, & Fabrigar, 1993). When two models are statistically equivalent, any observational dataset provides equal support for both of them; however, because they have different causal implications, it is possible to distinguish them via experimental intervention. Consider two statistically equivalent models: (I) A → B → C and (II) A ← B → C. Any possible observed correlation matrix between the variables A, B, and C will provide equal support for both models, though the arrow connecting A and B is reversed. The causal relation between A and B is therefore different for each model. Now suppose that A, B and C are measured at two points in time, and that between these time points, variable B is experimentally manipulated. Both the correlational structure at time point 1 and time point 2 will provide equal support for both models, but model (I) and (II) imply different expectations with respect to the increase or decrease in variable A as a result of the intervention on variable B. If intervening on B changes A to the extent predicted by model (II), that is evidence for model (II) over model (I).

Because causally different models can be statistically equivalent, some researchers and psychometricians argue that it is unjustified to make any causal interpretations of statistical models, and prefer the descriptivist approach, for which some reasons are given in Chapter 4. Though avoiding a causal interpretation in some situations is understandable, we would argue that, in certain contexts, interpreting and using the CFM as a causal model brings a number of benefits. We thus argue against the idea that the CFM is at all times merely a statistical model. In the next section, we take issue with one of the main arguments for interpreting the CFM as a statistical model rather than a causal model. According to this argument, a causal interpretation is not justified because factor models are typically estimated from the covariance structure of cross-sectional data without considering interventions on any of the variables over time.

5.3 Correlation Does Not Entail Causation

Factor analysis is typically applied to the correlation structure of cross-sectional data. However, a causal interpretation of the CFM implies that factor analysis infers causal relations between the latent variable and the indicators from observational data without having manipulated or intervened on anything. So how can factor analysis be used to hypothesize causal relations if it is based on correlational data?

Although correlations do not entail causality, causality does entail covariation. In addition, particular causal structures imply particular patterns of conditional independence among variables. For example, when A ← B → C is hypothesized, in which B is a common cause of A and C, then one expects covariation between A, B, and C, and one expects that A and C are conditionally independent given B (Reichenbach, 1956). Also, one expects that the correlation between A and C is smaller than the correlations between A and B and between B and C. These are all implications of a causal structure that can be verified from a correlation matrix. Similar expectations can be laid out for the reflective model (Figure 1), in which the latent variable is a common cause of its indicators: one expects the indicators to covary in a certain way, e.g., two indicators that are strongly correlated with each other should also share much variance with the other indicators. After all, indicators that consist of relatively more shared variance and less unique variance are more reliable indicators than those that consist of relatively little shared variance and more unique variance. Reliable indicators will also correlate more strongly with each other. Unlike in the 3-variable example cited earlier, however, it is impossible in latent variable modeling to test conditional independence given the common cause, because the common cause is unobserved. All in all, causation does lay out some expectations for observational data, though one should always consider the possibility of alternative causal structures that imply similar expectations for the data (e.g., the expectations summarized for A ← B → C also correspond to the structure A → B → C).

These implied constraints of the common cause structure are tested in the fit of a CFM. Violation of local independence will result in poor model fit. As noted earlier, although the hypothesis of a common cause can be rejected because of the testable constraints it puts on the data (e.g., on tetrad rank constraints see Bollen & Ting, 1993), the common cause structure cannot be verified from observational data alone because there will always exist alternative causal structures that can explain the observations in the data. But even though factor analysis cannot verify a common cause structure, hypothesizing such a structure justifies the use of factor analysis to measure the latent variable in the first place. After all, the constraints that are tested when fitting a CFM to the data are all implied by the common cause structure.

In the next section we elaborate on the descriptivist approach and then we proceed to developing three arguments in favour of a causal interpretation of the CFM.

5.4 What is the Common Factor in the Descriptivist Approach?

As described by Jonas and Markon (2016), the descriptivist approach entails that a common factor should not be interpreted as a real-world entity, but rather as mere shared variance that is nothing but a parsimonious summary of the data. From this perspective, latent variables are not postulated as concrete entities that have direct causal effects on their indicators, but rather as *summaries* of the covariance among indicators. When the latent variable is defined as the shared variance of a set of indicators, it cannot also be a common cause of the indicators. After all, the shared variance among indicators does not cause the shared variance among indicators; something cannot have caused itself. Thus, when the latent variable is the shared variance rather than being a cause of the shared variance, the arrows in Figure 1 pointing from the latent variable to the indicators should be interpreted as 'is part of'. The latent variable *is part of* the indicators, by the logic that the variance in the indicators *consists of* shared variance (arrow pointing from a common component to the indicator) and unique variance (arrow pointing from a unique component to the indicator). That is, when the latent variable is merely shared variance, the relation *intelligence* → *IQ test* should be interpreted in the same manner as *boys* → *people*: just as the set of boys is part of the set of people, intelligence is part of IQ test scores. Whether a variable is part of another variable or causes another variable, statistically speaking there is no difference between these two relations. More concretely, whether the latent variable is a part of the variance of the indicators (i.e., the variance that is shared among indicators) or is a cause of the indicators, both result in the same statistical model.

5.5 Toward a Causal Interpretation

Whereas descriptivists argue for a statistical reading of the factor model, we believe that a causal interpretation of the model is in many cases more appropriate. It is important to note here that our objections to the descriptivist approach are limited to cases in which the CFM is used as a measurement model rather than as a method for data reduction. When factor analysis is used as a data reduction method, the common factor can be used to predict other variables, but the common factor does not refer to anything outside the model itself. In data reduction, researchers are only interested in bringing a large set of variables down to a smaller set of dimensions. In other words, they are interested in a concise summary of the data, rather than in causal connections between latent variables and indicators. Again, in this case the obtained factor 'is part of' the original set of indicators, rather than referring to a common cause that generated the data, and a causal reading is not sensible.

In the next subsections we present three arguments for a causal interpretation of a CFM over a statistical interpretation when the goal is to measure psychological

constructs. We argue that (1) establishing causal relations conforms with most research questions in which the goal is to *explain* the correlations between indicators rather than merely summarizing them; (2) a causal interpretation of the CFM legitimizes *why* we are interested in the shared variance rather than in the unique variance of the indicators; and (3) a causal interpretation of the CFM legitimizes the assumption of local independence.

5.5.1 "Look, we found shared variance!"

Social scientists are typically interested in the best possible explanation for their observations. They want to know how and why observations occur the way they do. Borsboom et al. (2003) argued that latent variable models require a realist ontology in order to use them for establishing causal connections between latent variables and indicators. The choice for latent variable models, in which a set of indicators is assumed to covary *because* of this latent variable, already implies that one assumes this latent variable exists. Choosing a latent variable model therefore naturally implies a realist ontology (also see Hood, 2013). In other words, when the aim is to detect and theorize about a putative causal relation, it does not make sense to avoid a realist, causal interpretation all together.

To illustrate this point, we will use a study performed by Caspi et al. (2014) as a special case that in our view highlights why a causal reading is sensible. In their study, Caspi et al. (2014) "evaluate alternative hypotheses about the latent structure underlying 10 common mental disorders" (p. 120). The authors conclude that a bifactor model, with three group factors and one general factor, best explains the structure of psychopathology, and they name this general psychopathology factor the *p factor*. The descriptivist view would only allow the researchers to conclude that their study provides support for shared variance among a set of disorders. With a causal interpretation though, their findings would provide support for a meaningful hypothesis, namely that the *p factor* causes mental disorders to covary, and explains comorbidity. Here, a descriptivist approach is simply unsatisfying and uninteresting.

Of course it is legitimate to claim that a single dimension is able to account for most of the covariation among disorders, resulting in a parsimonious description of the data. But if the goal is to explain the structure of psychopathology, this claim is greatly unsatisfying: shared variance as such has no explanatory value. In contrast, when the shared variance reflects a common cause, this common cause does explain the correlations between indicators, resulting in a better understanding of the observations and an opportunity to further the research programme by searching for the identity of the common cause. A common cause not only renders the indicators independent when accounted for, the indicators correlate *because* they share a common cause. Explanation therefore has additional value to mere description, and in the end, establishing causal relations and finding explanations is essential to science. Thus, the *p factor* model has much

greater theoretical import as a causal model than as a statistical model. Researchers should, however, try to verify whether this causal model for the structure of psychopathology is justified in each case.

5.5.2 Why Shared Variance at all?

The CFM distinguishes between shared variance (that which is shared by all indicators) and unique variance (variance which is unique to each indicator). Shared variance is attributed to the latent variable, while unique variance is attributed to 'residual' influences, such as measurement error. The distinction between shared and unique variance, of which only the former is of interest in latent variable modeling, makes sense when a common cause underlies the indicators but is not sensible under alternative underlying causal structures. Put differently, the belief that a common cause underlies the indicators legitimizes that only shared variance is of interest rather than the unique variance of the indicators.

Consider an example in which the indicators reflect some variable that is not a common cause of the indicators. For example, *healthy eating* is a variable that can be indicated by responses to questions like "do you eat a lot of junk food?", "do you often eat fruit?", and "how much sugar do you eat?". These three items are all indicators of healthy eating, however, *healthy eating* is probably not a cause of these indicators. In this case it is not sensible to only take into account the variance that is *shared* by these items: the unique variance of the indicators is equally important to the construct *healthy eating*, even though these indicators are correlated and would likely result in a well-fitting reflective model.

The items *eating vegetables* and *avoiding sugar* are both relevant to the construct *healthy eating*, regardless of what causes them or how much of that variance is shared. That is, it does not matter *what* causes someone to eat a lot of vegetables (e.g., having a big vegetable garden) and whether that is the same cause as for not eating too much sugar (e.g., having strict parents). Both of these are indicators of healthy eating, including their unique variance. When a factor model is fit to these data, variance due to these unique causes is relegated to the uninteresting 'unique variance' component of the model. But arguably, the unique variance due to vegetable gardens and strict parents is no less relevant to *healthy eating* than the shared variance that may be due to causes such as a person's motivation to be healthy. As such, when the construct of interest is not a common cause, there is no theoretical reason to disentangle the shared and unique variance of the indicators. Doing so risks a biased representation of the construct of interest.

The above argument can just as well be applied to psychological constructs of which the exact nature is unclear. For example, *extraversion* is typically measured with a reflective model. Suppose that *extraversion* is just a summary of the scores on a set of extraversion items. In that case, why use the shared variance as a summary, rather than the unique

variance or all variance? What is the justification for discarding unique variance if the shared variance holds no special status? There is none. The shared variance is just one way of reducing high dimensional data to fewer dimensions, and there are alternative techniques for data reduction that do not discard unique variance, e.g., Principal Component Analysis (PCA; James, Witten, Hastie & Tibshirani, 2013). Thus, when the status of the construct is either known to be not a common cause (as in the *healthy eating* example) or is unknown (as in the *extraversion* example), the practice of interpreting only the shared variance among a set of items is not well supported.

In contrast, when one's theory explicitly states that the construct is a common cause of a set of items, it is immediately legitimate to interpret only the shared variance of the indicators because it is precisely this shared variance that must be due to the common cause. Returning to the Caspi et al. (2014) study, we can draw a similar conclusion. The reason why the authors are at all interested in the shared variance rather than the total variance, or the unique variance, is because they must have reasons to believe that the shared variance has certain explanatory power. Thus, the difference between a latent variable that is interpreted as merely shared variance and a latent variable that is believed to be a common cause of the indicators is that the former *results* from the distinction between unique and shared variance in a CFM, while the latter *legitimizes* this distinction.

5.5.3 Local Independence

The principle of local independence is a fundamental assumption of the reflective model and comprises the idea that when a common cause underlies a set of indicators, conditioning on this common cause renders the indicators statistically independent (Bollen, 2002; Borsboom, Mellenbergh, & van Heerden, 2003; Lazarsfeld, 1959; Lord & Novick, 1968). When factor analysis is performed on a dataset, it searches for a factor solution that meets this criterion. The assumption of local independence does not apply to all data reduction models. As stated before, PCA also results in a parsimonious summary of the data but in contrast to factor analysis, PCA does not use local independence as a criterion. PCA rather composes a variable that accounts for most of the variability in the indicators (James et al., 2013). In this process, PCA takes all variance into account, not only shared variance. Thus in the case of PCA, the indicators are not rendered independent, yet PCA provides a parsimonious summary of the data. A purely statistical interpretation of the reflective model does not legitimize the principle of local independence and therefore the use of the reflective model. In contrast, a causal interpretation of the reflective model implies that local independence should hold, so a latent variable model should be preferred when theory holds that a common cause is responsible for covariation among items. Getting back to the example of the p factor, the reason why the p factor is constructed in a way that it renders the disorders independent is because it is believed to reflect a psychopathology

factor that forms a common cause to these disorders. A mere summary of the data is not bound to such a constraint on the data. So not only does a causal theory justify a focus on the shared variance, it also justifies the assumption of local independence. Instead of local independence merely *resulting from* absorbing shared variance, a common cause *explains why* local independence is assumed.

5.6 How to Assess whether a Common Cause Structure is Correct

Using a CFM in the descriptivist framework does not require that a specific causal structure is hypothesized. It simply does not matter exactly what causal structure underlies the data for the latent variable to be a useful summary of the data. For the causal interpretation, in contrast, it is of central importance that the model accurately represents the underlying causal structure. Attaching causal meaning to a CFM that does not accurately represent the actual causal system will result in incorrect predictions.

Consider four indicators that all influence each other, resulting in shared variance among these variables. A CFM can be applied to the correlation matrix of these variables, and, depending on the strengths of the causal relations between the indicators, this may result in a well-fitting model. In such a situation, the factor model may be useful as a parsimonious prediction model, but as a causal model this factor model makes incorrect predictions with respect to intervention on any of the indicators. For example, a causal interpretation of the factor model implies that intervention on any of the indicators would not affect the other indicators, because all covariation is purported to arise due to a single latent common cause. In contrast, the true underlying causal structure implies that intervention on any of the indicators leads to changes in the other indicators, because the shared variance truly reflects direct causal relations among indicators. So a causal interpretation of an estimated model can lead to false conclusions when the causal structure of the data generating model is not represented accurately.

This begs the question of how to assess whether the true data generating model has a common cause structure. How do we figure out whether the principle of the common cause applies in cases in which we do not have easy access to the real underlying causal structure? As stated before, ideally one would manipulate the common cause, and see what happens to the variation in the indicators (Borsboom et al., 2003). If this were possible, one could directly observe the effects of the manipulation and consequently give causal meaning to the model. But when the common cause is latent, such direct manipulation can be impossible. In some cases, however, an alternative causal model might offer different predictions for manipulations on the indicators. For example, if an alternative model states that the indicators cause each other, such that the shared variance is explained by causal influences between the variables rather than by a common cause, intervening on the indicators would differentiate the two models. Whereas a model with causal influences

between indicators implies that certain indicators change as a result of interventions on other indicators, a common cause model implies that indicators are not affected by other indicators. These diverging predictions enable the researcher to differentiate between such alternative models that explain the shared variance among a set of variables.

When intervention is not possible, empirical tests of the goodness of fit of the reflective model may shed light on whether the common cause model is a likely generating model for a particular data set. Although a researcher cannot determine what causal structure underlies cross-sectional data, the common cause model does put testable constraints on the covariance structure of the data. Widely used test statistics and fit indices for confirmatory factor analysis test the hypothesis that the covariance structure of the data matches the covariance structure implied by a common cause model. Tests based on the pattern of partial correlations in the data may also allow researchers to determine whether a common cause model is more likely to underlie the data than an alternative model that posits direct effects among indicators (Van Bork, Rhemtulla, & Borsboom, 2015; Van Bork, Rhemtulla, Waldorp, & Borsboom 2016).

We believe that it is especially important to think about the data generating process in the phase of test construction. As we mentioned before, researchers do not pick an arbitrary set of indicators but select those indicators that are hypothesized to be affected by the construct of interest. This is an important point that concerns the stage of test construction rather than test-analysis. Borsboom, Mellenbergh, and van Heerden (2004) write:

> a century of experience with test construction and analysis clearly shows that it is very hard to find out where the scores are coming from if tests are not constructed on the basis of a theory of item response processes in the first place. (p. 1067)

They continue with the conclusion:

> Thus, it is suggested here that the issue may not be first to measure and then to find out what it is that is being measured but rather that the process must run the other way. It does seem that if one knows exactly what one intends to measure, then one will probably know how to measure it, and little if any validation research will be necessary. (p. 1067)

Another possibility to give causal meaning to the factor model, is that a factor model can be posited as a plausible hypothesis for a certain phenomenon (Haig, 2005a; 2005b; 2014). When theoretical considerations imply that the principle of the common cause

describes a certain mechanism accurately, the factor model can be used as a method for theory generation. Specifically, exploratory factor analysis can be used for the generation of plausible theories, and through confirmatory factor analysis, these theories can then be evaluated. Through methods such as cross-validation and replication, the theory can gain additional support. This of course does not result in absolute certainty that the causal reading is in fact the true reading, but by choosing the models that have more explanatory value over those that are less explanatory, the best model is left standing. This way, it is possible to gather evidence for the hypothesis that a common cause structure is indeed applicable to the phenomenon in question.

All in all, an inevitable conclusion of the causal interpretation of the factor model we defend, is that the data generating process matters. Therefore, one should always try to assess whether it is plausible that the data generating mechanism has a common cause structure.

5.7 Conclusion

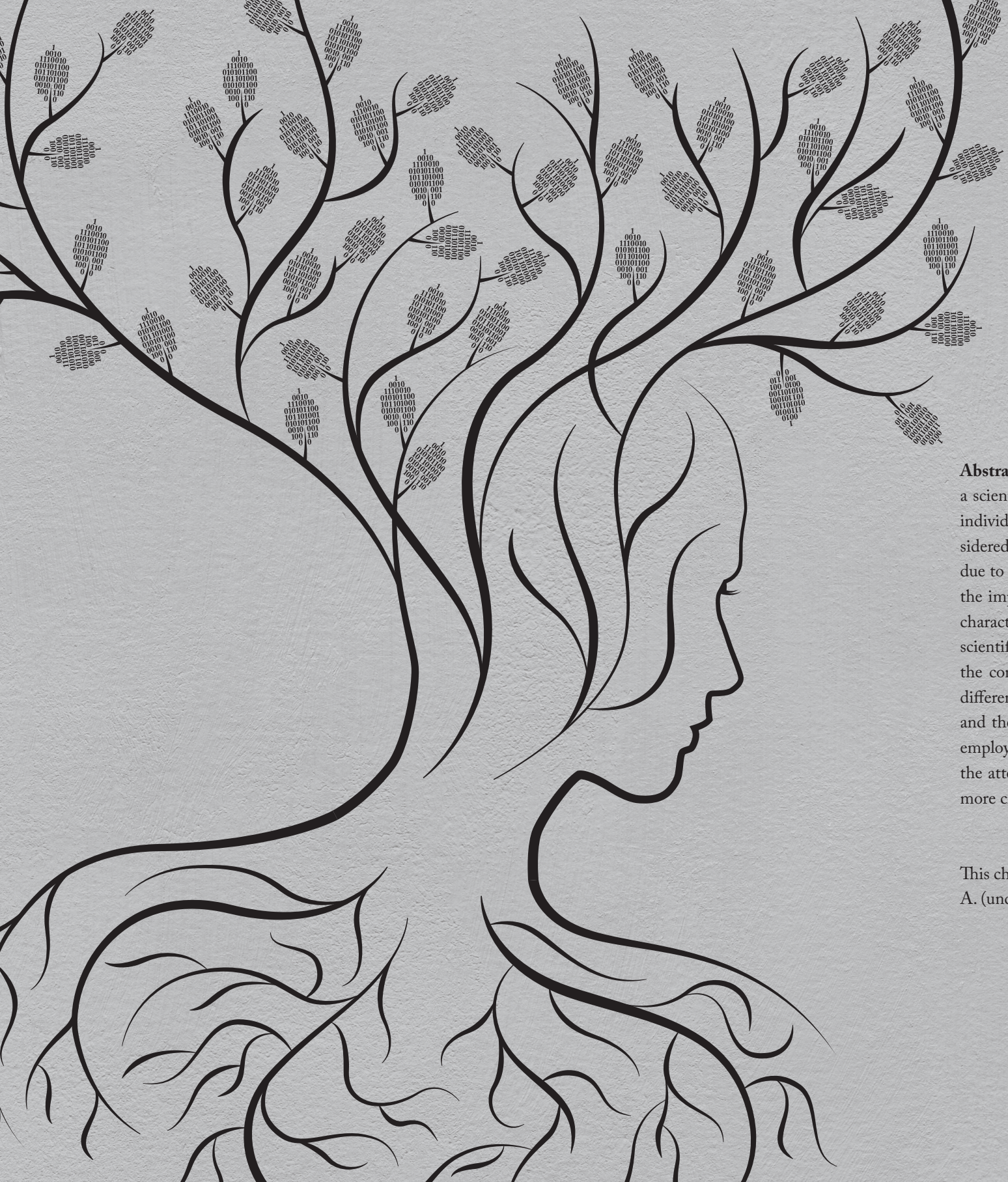
In this chapter we argued that, if the goal is to measure psychological constructs and make meaningful claims about how such constructs relate to each other, a causal interpretation of the CFM rather than a merely statistical interpretation proposed by the descriptivist approach is preferred. First, a causal interpretation matches the scientific aim of explaining observed patterns in behavior rather than summarizing them. Second, a causal interpretation of the CFM legitimizes *why* the shared variance is of interest rather than the unique or the total variance of the indicators. And lastly, a causal interpretation of the CFM legitimizes the assumption of local independence. A statistical reading of the model is not incorrect but does not provide an explanation for the observed data, nor does it explain why shared variance rather than unique variance is of interest and or explain why local independence should hold.

We are aware that a causal interpretation of reflective models brings a host of other problems. Reading a reflective model causally does not suddenly turn a non-causal mechanism into one that is in fact causal. Reflective models are only tools to posit plausible causal relations, and they need additional input in the form of causal knowledge or assumptions to establish what kind of causal relation it is. In other words, researchers construct plausible causal hypotheses, fit a model that imposes these hypotheses on the data, and then consider how the output supports their hypothesis. This process has a built-in buffer that balances formulating strong causal statements that refer to real-world mechanisms and being cautious about such statements.

We also acknowledge that there are situations in which a statistical non-causal interpretation of the reflective model makes more sense than a causal interpretation. When a common cause structure is unlikely to underlie the data, a purely statistical

interpretation of the reflective model can still be useful for prediction or data reduction. For example, although the existence of general intelligence is often disputed, the *g*-factor can be used to merely predict school success, without having to establish its objective existence. As elaborated on in Chapter 4 of this dissertation, positing causal claims is an interpretation of choice and there are other purposes of psychometric models that are not of lesser value than a causal interpretation.

However, as we argued in this chapter, a causal interpretation of the CFM has certain benefits, and for the sake of theory building, we would plead for an understanding of the CFM as an explicit causal hypothesis that can be falsified. In situations where the data generating mechanism matters for the theory about a psychological construct (e.g., should depression be understood as a cause of its symptoms or as the name for a system of interacting symptoms?), treating the CFM as a summary rather than a hypothesis about this causal structure takes away the need to test the data-generating mechanism. Rather than concealing the urge to infer causal relations by restricting the interpretations of these models to descriptions of the data without reference outside of the data, we argue that one should dare to hypothesize. By making hypotheses explicit (e.g., "we hypothesize that construct A is the common cause of the indicators X_1 to X_j"), they are open to falsification. Additionally, interpreting the CFM causally stresses the need for theories about *how* differences in the latent variable result in differences in the responses; theories that we think are crucial for the measurement of psychological constructs.



Chapter 6

Values in Psychometrics

Abstract When it originated in the late 19th century, psychometrics was a field with both a scientific and a social mission: psychometrics provided new methods for research into individual differences, and at the same time, these psychometric instruments were considered a means to create a new social order. In contrast, contemporary psychometrics – due to its highly technical nature and its detachment from applied research – has created the impression of being a value-free discipline. In this article, we develop a contrasting characterization of contemporary psychometrics as a socially and politically value-laden scientific discipline. The values in contemporary psychometrics that we elaborate on are the conceptualization of individual differences as quantitative (rather than qualitative) differences, the aim for objective measurement, the formalization of fairness of test items, and the preference of utility above truth. Our goal is not to criticize psychometrics for employing social values, but rather to bring the values in contemporary psychometrics to the attention of both applied researchers and psychometricians, and to invite them to a more critical and contextualized reflection of the field.

This chapter is submitted for publication as Wijsen, L. D., Borsboom, D., & Alexandrova, A. (under review). Values in Psychometrics. Submitted to *Perspectives in Psychometrics*.

6.1 Introduction

Early psychometrics – the discipline that is concerned with the measurement and prediction of human traits, aptitudes, and behavior – was deeply invested in social and political developments, such as the eugenics movement, the introduction of military testing during the world wars, and the rise of a national education. As a result, its methods and tools were explicitly geared towards furthering the goals of these movements. Today, however, psychometrics is a conventional academic discipline whose primary goals appear to be mainly scientific, making it harder to detect any deeper political or social allegiances. But they nonetheless exist and in this paper, we draw on recent developments in history and philosophy of science to develop a characterization of psychometrics as a socially and politically value-laden scientific discipline.

When we speak of psychometrics in this paper, we do not aim at psychometrics in all its shapes and forms. Rather, we speak of psychometrics in relatively narrow terms: the part of psychometrics that is (almost) strictly model-based. The psychometricians we target in this paper tend to work on the technical aspects of psychometric models, such as Item Response or Structural Equation models. Models of this kind often explain the relationship between a latent variable that represents a psychological construct, and the items that this variable supposedly measures. To draw a sharp contrast here: our target psychometricians mostly work on developing the technical aspects of these models (e.g. extensions, methods for parameter estimation, methods for model fit assessment), whereas many other psychologists and also psychometricians apply these models to understanding a construct of interest (e.g., cognitive abilities, personality dimensions, or psychological disorders) and its relation to other attributes or aspects of society. The psychometricians we discuss here are thus of the first kind, not the psychometricians with specific substantive interests who use psychometric models to investigate these interests. The Psychometric Society is an institution that is most representative of our target field and *Psychometrika*, its flagship journal, publishes the research we are aiming at: technical or theoretical papers that elaborate on the technical or statistical features of psychometric models. Thus, the psychometricians we aim at in this paper *develop* psychometric models or new aspects of these models, rather than investigate specific substantive questions about a psychological topic for which they use psychometric models. For the sake of clarity, we will denote this subfield as 'psychometrics' in the remainder of this article, but we acknowledge that not everyone who identifies themselves as a psychometrician shares this strictly technical approach.

The practice of psychometrics, whether theoretical or applied, requires its practitioners to make value judgements, and we thus speak of a psychometrics that is value-laden. More specifically we use 'value-laden' to pick out two features: a) these value judgments concern morality and social life and not just epistemic goods such as truth and

empirical adequacy, and b) these value judgments enter into the appraisal of psychometric models and hypotheses, not just prior decisions about what to research, nor how to apply existing knowledge. Consequently when we speak of psychometrics as having a 'value-free' image we mean the denial or the failure to recognise (a) and (b) above. Value judgements then are explicit claims or implicit commitments that animate scientific inquiry and that enable inference and theorising where empirical or logical considerations do not uniquely compel any answer. Some value judgments are claims about what inference or theory is most predictive, most explanatory, or is empirically adequate; these are known as epistemic values (Douglas, 2000, 2009; Longino, 1996; Rooney, 1992). Value judgments about what is considered *good*, *right*, *just*, or *beautiful* are known as non-epistemic values. Recent philosophy of science recognises a variety of ways in which epistemic and non-epistemic value judgments enter both pure and applied scientific research at all stages, from research planning to hypothesis testing (Douglas, 2009; Elliott, 2017).

Values in psychological science have been addressed on a number of occasions, and in different shapes and forms. For example, some studies investigate the need for diversity of political preferences in academic psychology (Duarte et al., 2015; Redding, 2001), other studies investigate how psychology can become a more socially engaged, or socially conscious, discipline (Gergen, 1973; Nafstad & Blakar, 2012). Closely related to this paper, are the several studies that address the role of values in the process of educational measurement (Gordon & Terrell, 1981; Messick, 1975, 1989; Stein, 2014). However, the role of value judgments in contemporary psychometrics, and specifically how they are embedded in much of the technical aspects of psychometric research, has not been explored in the literature so far and is our focus in this article.

We show that there has been a gradual shift in the 20th century from a psychometrics that is explicitly socially engaged and eagerly participating in the project of social improvement to a contemporary psychometrics that sees itself as a purely technical discipline in search of formal tools that may later be used for testing and ranking. We argue that, although social values are not as readily seen in today's psychometrics and although the discipline goes to great lengths to cultivate the image of an abstract pursuit in search of neutral quantitative tools, it is by no means a value-free discipline. We discuss four senses of value-ladenness in contemporary psychometrics: the conceptualization of individual differences as quantitative (not qualitative), the aim for objective measurement, the aim for fair measurement, and the preference for utility above truth. When commitments such as objectivity are endorsed, they are not evidence of value-freedom. Instead, they are evidence of the endorsement of a specific social value: seeking to remove personal judgment – as far as possible – from the testing process and to substitute it with a combination of good assessment procedures and proper data analysis. In doing so, today's methods of modeling implicitly endorse the ideal of a meritocratic society where people can be ranked on the

basis of certain socially valuable measurable characteristics pertaining to their abilities, and ranked in a way that can be justified on purely procedural grounds.

In articulating the ways in which even the most technical work in psychometrics takes moral stances, our goal is not to criticise the discipline, nor to expose it as somehow biased and failing to live up to ideals of science. Any inquiry needs to make the sort of foundational bets we attribute to psychometrics, for without them the scientific work cannot get started. However, exactly what commitments a discipline adopts are neither inevitable, nor innocent. Making them explicit will hopefully contribute to a more reflective psychometrics and will enable psychometricians to judge the extent to which their work as scientists makes them responsible for subsequent applications of psychometrics in the wider world.

6.2 A Brief Note on Terminology

In this paper, we define psychometrics narrowly and our analysis thus pertains to psychometrics narrowly defined. To speak of values in psychometrics as a whole would require a wide array of evidence (e.g. interviews, survey data, citation analysis), the scope of which goes beyond the scope of this paper. Our contribution however adopts a methodology common in philosophy of science, in which normative commitments of a field are inferred from the implicit commitments in some of its most central research projects. There are of course limits to this methodology, since it does not pick up on the attitudes of psychometricians that cannot be deduced from their publications, nor can we speak for the entire scope of psychometrics. Nevertheless, because we show that values are already detectable at this technical level of psychometric research, we expect that they are also detectable at different levels of technological advancement, especially when using a variety of research methods such as the ones mentioned above, and we hope that our research might inspire more inquiries into this topic in the future.

We make a distinction between implicit and explicit expressions of values, which also needs some additional elaboration. Values can be implicit in that they have not yet been formulated in other studies – philosophical or otherwise – before. More importantly, values can be implicit in that they can be hidden under formalisms, which is often the case in contemporary psychometrics. Values in psychometrics are usually deployed as entirely technical decisions or considerations, whereas our argument is that they are not just technical; they carry with them a moral component. To the extent as they appear, they do not appear as values. For example, fairness as a term occurs in many psychometric research papers, but not as a value per se: the idea of fairness occurs mainly as a formalization of an idea and is captured in psychometric jargon or statistical equations. However, there is still a moral component here that is substantially different from other types of fairness (which we will discuss later). The value of fairness is thus implicitly rather than explicitly embedded in psychometric research.

6.3 Values in Early Psychometrics

Before we continue with our analysis of values in contemporary psychometrics, it is necessary to take a brief look at values in early psychometrics. Our aim is to set up a contrast with contemporary psychometrics, by showing that early psychometrics was explicitly intertwined with important political and social movements of the time, and thus with specific social values. Below we elaborate on two of these movements, the ideology of eugenics and the rise of military and standardized testing, and show how the values that were part of these movements were mirrored in early psychometric research. Our aim here is to not to speak about psychometrics as a whole, but make a comparison between how historical gatekeepers were engaged in political ideas, and how current gatekeepers, though sometimes committed to values we will discuss later, shy away from making large political statements.

6.3.1 Eugenics and a New Social Order

One of the ideologies in the late 19th century that was crucial for the emergence of psychometrics as a scientific discipline was eugenics; the ideology that a population could be and should be improved by encouraging people with desirable traits to procreate, and discouraging people with less desirable traits to do so. Several early psychometricians were active eugenicists, which in the late 19th century and early 20th century was not a very controversial ideology to support. Examples of such scholars were psychologists James McKeen Cattell (1890), Cyril Burt (1909, 1955, 1957), Charles Spearman (1904, 1927), and bio-statistician Karl Pearson, the first editor of *The Annals of Eugenics* from 1925 to 1934. Eugenics was considered a tool to create a 'perfect' society, where people with the most talent would end up in the most important positions, and people with less talent were assigned the easier, less esteemed, tasks. Ideally, the human race could be improved or perfected through selective breeding.

Francis Galton, the inventor of eugenics, stood at the birthplace of psychometrics as a discipline (Buss, 1976). Galton, on the one hand, aimed for the freedom for people (or men) to develop and specialize in a wide variety of directions, regardless of the class in which they were born, but on the other hand, believed that the order of society should be based on human ability. Given that everyone had the opportunity to develop their own abilities, Galton believed that the hierarchical class structure in British society at the time must represent the differences between abilities. Galton considered measured abilities, with intelligence as its prime example, to be hereditary qualities (Galton, 1869), and though he did not deny the possibilities of external influences such as training and education, he believed there was an innate, person-specific, limit to people's abilities (White, 2006). Galton was one of the first to take up the task to measure the differences between people using quantitative methods.

In the history of psychometrics, Charles Spearman is considered one of the most influential thinkers, for he was the first to use the quantitative Galtonian and Pearsonian techniques, such as correlation and regression, for the formalization of the relationship between psychological attributes, such as cognitive abilities, and the test items that are supposed to measure this attribute. Spearman (1904) believed he had found empirical evidence that the reason why certain people score high on tests, whereas others do not is because of their level of *general intelligence* or *g*. He considered general intelligence a latent variable, most likely in the shape of mental energy, that has a direct causal effect on how people perform on tests. Spearman thus considered *general intelligence* as a common cause of the observed variables. The statistical model in which Spearman formalized these ideas has become known as the common factor model. In Spearman's common factor model, the latent variable or common cause represented general intelligence which was measured through a set of ability tests, but throughout the years, latent variables have denoted all sorts of psychological attributes like personality dimensions, cognitive skills, and psychiatric disorders. This model was the first latent variable model and inspired the development of many others in the 20th and 21st century.

Spearman had a strong scientific curiosity to uncover the structure of the mind by using objective, statistical methods. These methods had turned out successful in the natural sciences, and were now bound to turn the social sciences into an objective and rigorous endeavour. His mission was also to contribute to "psychology of a more exact character" (Spearman, 1914, p. 25), in which abilities can be "definitely measured and permanently recorded", rather than being estimated by "hearsay, causal experience, and remote reminiscence". Spearman was a member of the Eugenic Society and considered general intelligence to be standing at the basis of social institutions (Spearman, 1914; 1927). Spearman's common factor model was a good match with the politics of eugenics: Spearman considered general intelligence to be a hereditary quality that resided in the brain (though he was unsure where exactly), and the common factor model could contribute to measuring people's intelligence in an objective fashion and to shaping a society that was ordered based on this quality. Compared to other eugenicist psychologists at the time, Spearman was perhaps not as vocal about his eugenic ideology and quotes like "Every normal man, woman, and child is, then, a genius at something as well as an idiot at something" (1925, p. 439) suggest that he also believed that one can be a genius in many aspects, not just when one excels at general intelligence. However, it would not be a stretch to say that he considered the measurement of general intelligence to be crucial to solve certain social problems, such as admission into citizenship and even "for the right of having offspring" (Spearman, 1927; p. 8).

6.3.2 Military Testing as Spin-Off for Standardized Mental Testing

One of the most important instigators of widespread mental testing, especially in the United States, was World War I. During World War I, a total of 1,75 million American soldiers were tested on intelligence and emotional functionality. Based on the test scores, decisions were made about the appropriate position for the prospective soldier within the army, and sometimes the decision led to a disqualification of the prospective soldier from the army entirely. The success of military testing was seen as proof of the merit of mental testing as a method to set up a social order; the success of testing in the army was evidence that tests were "legitimate means of making decisions about the aptitudes and achievements of normal people - an essential means for making objective judgments about individuals in a mass society."(Reed, 1987, p. 76). Shortly after the war, mental testing entered the educational system and was used to evaluate the abilities of millions of school-children. According to psychologist and eugenicist Lewis Terman – who introduced the first intelligence test for American children in 1916 (the Stanford-Binet Intelligence Scales, a revision of the originally French Binet-Simon scale) – mental testing served the American meritocratic democratic ideal. The welfare of American society depended on the education of people with high innate ability, and mental tests were needed to ensure that people with high abilities could be identified, so that they would receive the proper education and at one point, could lead the nation (Minton, 1987). Tests were also expected to "bring tens of thousands of these high-grade defectives under the surveillance and protection of society (Terman, 1916)", so that crime and poverty could be eliminated. Mental testing was thus considered the most effective method to reach a meritocratic social order. In the words of Minton: "'Prediction and control', 'human engineering' and 'social efficiency' were the catchphrases for postwar American psychology." (1987, p. 106).

Another example of the social involvement of early psychometricians like Lewis Terman and Alfred Binet, but also early psychologists like William James and Stanley Hall, was their devotion to the practical side of education, such as the methods of teaching and problems in the classroom. They wanted to investigate how science could contribute to solving the practical problems that teachers and students come across. However, from E. L. Thorndike onwards (one of the first presidents of the Psychometric Society and one of the founders of educational psychology), educational psychology became less and less involved with the practice of education (Berliner, 1993). Under Thorndike's leadership and his dedication to reinforcement theory, educational psychology became largely laboratory-based and separated from the classroom. Educational psychology's connection to the practice of education, and thus its social involvement, became weaker under Thorndike's influence.

In the 1940s, eugenics fell out of grace due to the rise of Nazi politics and practices, but the popularity of (standardized) testing only increased. By the 1930s, testing had

become normal practice in society, and "presented itself as a technocracy vital to a society which was generally becoming more test-conscious and meritocratic." (Evans & Waites, 1981, p. 25). National education systems were in place, and testing fulfilled the bureaucratic needs of these large infrastructures. Psychometrics had the opportunity to blossom. In 1937, Louis L. Thurstone founded the Psychometric Society, which remained for the most part of the 20th century a mostly North American institution. Though eugenics was still in full swing at the time, the Psychometric Society did not have an explicit eugenicist motivation (or any political motivation for that matter): it rather aimed to support and encourage strong quantitative 'rational' research in the behavioral sciences, often (though not exclusively so) with an emphasis on measurement.

The social and political developments discussed above and our claims on their mingling with early psychometrics have been widely accepted by historians of early psychometrics, and are in that sense not controversial. However, having emphasized the importance of these developments to early psychometrics, we are now in the position to show that there has been a gradual shift towards a psychometrics that is less involved with politics, and also less involved with psychological science.

6.3.3 A Gradual Shift in the Self-Conception of Psychometrics

As we have seen above, early psychometrics was strongly intertwined with social and political values. Psychometricians like Francis Galton, James McKeen Cattell, and Lewis Terman endorsed a socially involved vision of their discipline. According to this vision, psychometrics would contribute to a new, meritocratic social order based on intelligence, which was often considered to be part of a person's biology and thus heritable (Gould, 1981). But at the same time, they shared a scientific curiosity about the possibilities of quantitatively measuring people's hereditary traits, and they considered proper measurement instruments capable of doing that in an objective fashion. So, it was quantitative measurement that enabled the possibility to distinguish between people with different levels of intelligence or other abilities, and to distinguish between for example the gifted students and the weak students (or 'the feebleminded', e.g., Terman, 1916).

After World War II, the eugenics movement had lost its social approval (though, as Stern (2005) argues, eugenic thinking remained influential until the 1970s). Mental testing, on the other hand, proved an effective method for the evaluation or diagnosis of people's intelligence, aptitudes, and capacities; through tests, large quantities of people could be measured simultaneously. The measurement and determination of individual differences was also considered important for society at large: it enabled the precise design of a society in which people with high intelligence took in highly specialized positions. Even though the explicit eugenic influences of psychometrics' early days faded over time, mental measurement continued to remain important in several areas of society, such as

education. Mental testing thus became part of social reality and has since pervaded many lives (as it still does). The success of testing resulted in vast amounts of data that were in need of reliable analysis, providing a fruitful base for psychometric research to flourish. Gradually, the conditions were laid down for psychometrics to become less and less involved with explicit political and social motives.

Besides a gradual erasure of an explicit commitment to social and political values in psychometrics, such as those visible in the work of the early eugenicists, there was also a gradual movement away from psychology itself. Besides being committed to social or political motives, early psychometric research was also often committed to turning psychology into a quantitative, reliable science, and psychometricians developed methods or models to uncover specific psychological mechanisms. Among other topics of interests, Charles Spearman was dedicated to uncovering the laws of general intelligence (1904), James McKeen Cattell to the measurement of separate functions like memory span and reaction time (1890), and Louis L. Thurstone to designing measurement scales for the measurement of attitudes (1928).

However, during the second half of the 20th century, psychometrics became more and more detached from psychological theory (Borsboom, 2006; Sijtsma, 2006; Wicherts, 2007). In the decades that followed the Second World War, psychometricians developed a number of statistical modeling techniques which then became the dominant research traditions in the field: Item Response Theory (IRT, see Hambleton, Swaminathan, & Rogers, 1991), Structural Equation Modeling (SEM, see Jöreskog, 1970, 1973), and Multidimensional Scaling (MDS, see Kruskal & Wish, 1978). More so than in the early 20th-century psychometrics, these modeling traditions, and especially their many extensions that were developed in the years that followed, were not necessarily developed to build a theory about psychological mechanisms. They either aimed at solving a data-analytic or inferential problem (which is a central feature of techniques such as SEM and MDS), or at proving a more reliable method for estimating people's abilities and item difficulties (IRT). A good example of the distinction between a psychometric model on the one hand, and a substantive theory on the other, is given by SEM. The development of SEM enabled the modeling of the structural relationships between a set of latent variables. The technique as such assigns no meaning to the variables; the model itself only provided a method for analyzing relationships between multiple (possibly latent) variables. Assigning substantive meaning to the latent variables is the task of the psychologist who uses the model for theory construction.

Over time, psychometrics turned into a largely model-based discipline, more embedded in statistics than in psychology, which could be applied to different topics within psychology and other disciplines. As such, psychometrics was no longer affiliated with intelligence alone, but applied in a wide variety of subfields in psychology. As we

argue in the next section, these developments created a new image of psychometrics that, on the face of it, is free of applications, free of psychological content, and consequently free of values.

6.4 'Value-free' State-of-the-Art Psychometrics

As we described earlier, the kind of psychometrics we aim at here is a very abstract and technical kind, that mostly deals with measurement models. Though psychometrics has incorporated statistics and techniques from outside the field, (extensions to) IRT and SEM are still two of its main research focus areas. Examples of important projects in contemporary psychometrics are Computerized Adaptive Testing – a form of computer-based testing which adapts the items to the examinee's ability level –, cognitive diagnosis, Bayesian estimation of psychometric models, methods for model evaluation and comparison, and response time analysis. *Psychometrika* is the flagship journal of The Psychometric Society, and psychometrics' flagship journal. Articles in *Psychometrika* are known to be fiendishly technical and difficult to understand for researchers who lack a strong statistical or psychometric education. Only a small community of statistically trained psychometricians is able to read and review most psychometric literature. Though there have been efforts in the past to run a successful applied section in *Psychometrika*, *Psychometrika* has remained a mainly theoretical and technical journal. Its content is increasingly instrumentalist and 'content-empty'; latent variables, once considered innate 'measurable' entities, are now often considered random effects, convenient for the process of modeling and estimation. The emphasis lies on data analysis of behavioral data, rather than the measurement of substantive psychological constructs. In contrast, *Psychological Methods*, a journal that publishes a wide variety of article types on quantitative methods in psychology (among which psychometric methods, but not exclusively so), explicitly encourages authors to make their work understandable to applied researchers. Such an injunction might not have been necessary in the past, but it is necessary now that most psychometric literature is unreadable for most applied researchers.

Psychologists and other applied researchers, often more interested in substantive problems, may be inclined to leave the highly technical psychometric literature for what it is, and the psychometricians are not always sensitive to the problems applied researchers deal with (Sijtsma, 2009; Young, 1996). This reciprocal lack of involvement with the other party contributed to the status of psychometrics as a field without a very strong connection to substantive psychological research. The changes from psychometrics as a substantive research area with a strong connection to psychology to a highly abstract and technical field has made the detection of an explicit political agenda in this research difficult. Contemporary psychometric research - e.g., research on the technical properties of IRT models or a Bayesian method for model comparison - invites the impression of

being a field with no apparent social mission and with only technical tools to offer. This impression motivates the idea that contemporary psychometrics is value-free or at least more so than it was formerly.

Another manifestation of psychometrics' apparent value-freedom is shown by the departure of validity from the psychometric jargon. 'Validity' became a frequently used term in psychometric literature in the 1910s and 1920s (Newton & Shaw, 2014), and was conceptualized by psychometrician Truman Kelley (1927) as "whether a test really measures what it purports to measure" (p. 14). Witnessing the rise of intelligence tests, he found it important to be careful about drawing bold conclusions about a person's attributes, and argued that it was first and foremost important to know whether a test indeed measures what it aims to measure before one draws these conclusions. The definition of validity has gone through a number of changes, and contemporary definitions emphasize that validity is a matter of sufficient evidence that supports specific interpretations and uses of a test (Kane, 2001).

Though psychometricians still *care* about validity in a general sense – they would agree that it is of vital importance that measures are thoroughly checked for validity –, validity is no longer a common research topic in contemporary psychometrics. So even though validity is possibly one of the best-known and most exported concepts in psychometrics, and still a value for the psychometricians themselves, the discussion of what validity entails and how to establish it has largely moved away from psychometrics proper. Instead, this debate takes place in education and psychological research (e.g., Moss, 1995; Kane, 2001; Newton & Shaw, 2014). The departure of validity from psychometric discourse is a sign of psychometrics' increasingly instrumentalist approach and value-free self-conception.

Another aspect of psychometrics' perceived value-freedom resides in the fact that it does not endorse particular uses of testing for social purposes and does not recommend any particular educational or workplace policies. The debate on the role of tests in society is part of public discourse both in the United States and the Netherlands (two of psychometrics' strongholds), but is mostly led by policymakers and education reformers (not by the psychometricians). According to Evans and Waites (1981), psychometrics has never been pure but applied knowledge, endorsing the ideal of a 'technocratic rationality' which commands that any problems be solved within the framework of the technology itself. Seen as technocratic rationality, psychometric research is defined as a series of mostly technical and statistical problems that need to be solved, and does not concern itself with long term goals in terms of desirable social, economic, and political development. Because psychometrics is defined as a series of technological problems, psychometrics presents the social and political interests that flow 'naturally' from the technology as value-free.

The apparent value-freedom of contemporary psychometrics thus follows from psychometrics' highly technical content, its somewhat isolated existence from applied research, and its tendency to operate within the constraints of technocratic rationality. In what follows, we argue that this resistance to issuing specific social recommendations does not amount to value-freedom and that values still nevertheless permeate and inform the practices of contemporary psychometrics.

6.5 Values in Contemporary Psychometrics

Since standardized testing is always a normative endeavor – it sets benchmarks and categorizes people as meeting or failing to meet these benchmarks – the process of testing and the instruments are strongly laden with judgments about what is worthy, proper, or suitable (Stein, 2014). Moral values are also strongly visible in what we consider 'pre-psychometrics'; the process of deciding which abilities and aptitudes are considered important and in need of measurement (a process which lies more in the hands of policymakers and applied psychologists than in the hands of psychometricians themselves). Values in the realm of testing, such as the ideas of social justice or equal opportunity, have been discussed by several authors (Gordon & Terrell, 1981; Messick, 1975, 1989a; Stein, 2014), and remains a topic worthy of further investigation. What we will discuss in this article however, are values in psychometrics proper, a topic not previously addressed.

In the following, we make a distinction between four values that permeate psychometric decision making in different ways: the conceptualization of individual differences as quantitative (not qualitative) differences, the aim for objective measurement, the formalization of fairness of items, and the preference of utility above truth. Though we treat these values separately, there are areas of overlap, which we will highlight as we go through the analysis.

6.5.1 Individual Differences are Quantitative, not Qualitative

Psychometrics is committed to investigating *individual differences* in a quantitative fashion. Individual differences are typically considered to be naturally ordered, or even quantitative - that is, if two individuals differ in intelligence or a scholastic aptitude like reading ability, one of them is assumed to have more of it than the other, rather than having a different *kind* of ability. Psychometric models, and especially Rasch models, clearly reflect this assumption (Bond & Fox, 2015), and this commitment has not significantly changed over time. In fact, for the most part, psychometrics appears still dedicated to developing models for quantitative measurement. We argue that the choice for conceptualizing individual differences as quantitative, not qualitative differences, involves two values: the epistemic value that only quantitative knowledge counts as properly scientific, and a moral

commitment to a specific form of equality, namely that differences between human beings are only of degree and not of kind. Let us illustrate each in turn.

The imperative that proper science is quantitative is widely known to have influenced early psychology (Michell, 2003). This is clearly visible in the work of early psychometricians. For example, Francis Galton believed that the only means to a scientific approach to the study of the mind is the measurement of quantitative differences (Galton, 1879). Quantitative research had proven successful in the natural sciences, and psychometricians were confident that it would also become the method to investigate laws of the human mind. McKeen Cattell's first two sentences in his 1890 article illustrate this ardent belief:

> Psychology cannot attain the certainty and exactness of the physical sciences unless it rests on a foundation of experiments and measurement. A step in this direction could be made by applying a series of mental tests and measurements to a large number of individuals. (p. 373)

The view that a quantitative approach makes research more rigorous and reliable, and thus more scientific, is still widely held by psychometricians. By conceptualizing abilities and aptitudes as quantitative attributes, and by developing quantitative methods to analyze and measure these abilities, both early psychometrics and contemporary psychometrics hoped to draw reliable conclusions about individual differences.

There have been a number of more qualitatively oriented approaches in psychometrics worth mentioning. One example is Thurstone's theory of primary mental abilities, which assumes that intelligence does not consist of one factor, but multiple independent factors (Thurstone, 1938). Though people differ on these factors in terms of degree, there is also qualitative differentiation: everyone has an individual 'profile' of levels on the different factors. A second example is Aptitude x Treatment interaction which investigates how different methods of instruction interact with student ability and achievement (Cronbach & Webb, 1975). There has also been the occasional psychometric model for qualitative differences, which assumes a discrete, rather than continuous, latent variable (e.g., the latent class model). A class in such a model accounts for a specific pattern of item responses. These models however are not mainstream in psychometrics. Moreover, and perhaps unsurprisingly, the latent class model originates in the work of sociologists Lazarsfeld and Henry (1968), not in psychometrics, and Cronbach & Webb (1975) was published in the more theory-oriented *Journal of Educational Psychology,* not in the more technical *Psychometrika*. All in all, it is safe to state that psychometrics relies heavily on a quantitative approach.

In psychometrics, the commitment to quantity is not one justified to everyone's satisfaction. There is no proof that attributes are in fact quantitative in nature (Michell, 1999), and as previously mentioned, contemporary psychometricians do not often entertain the belief that an attribute exists in reality. But this lack of certainty does not hurt the case for quantification because, in addition to being a commitment about proper scientific method, it also has a moral dimension – in particular, the commitment to quantification can be seen as a commitment to a form of equality. This may sound ironic since psychometrics is the basis for discrimination and rankings, which result in unequal outcomes for all involved. Besides, it is very hard to attribute a commitment to equality to the early practitioners of psychometrics who were in search of a superior race and a superior gene pool. This irony notwithstanding, today's psychometricians adhere to a particular type of equality while at the same time rejecting the embarrassing allegiances of their predecessors. The sense of equality we have in mind is equality of the standards against which all individuals will be judged, the standard often being a scale or distribution on which individuals take different positions. To quantify individual differences as psychometricians have done implies a strong value judgment about how people *ought to be* evaluated: namely, only on that specific trait. This is both different from equality of opportunity (though, as we shall see shortly, this commitment is also important to contemporary psychometrics) and different from a commitment to equality of resources or endowments (as discussed in Dworkin, 1981).

In addition to the moral and the epistemic justifications for quantification, there is also a deeply practical one: numerical rankings enable psychometricians to give pithy summaries of enormous amounts of data and information about individuals. As we shall see below, the commitment to being 'practical' is another crucial commitment of contemporary psychometrics. Comparing people's abilities on a single quantitative scale is highly practical and more effective than if we had conceptualized differences between people qualitatively. Decisions based on qualitative data are typically hard to trace, not very transparent, and as such do not easily lend themselves to psychometric analysis. This aligns with Porter's well-known argument that the quantification "minimizes the need for intimate knowledge and personal trust" (1995, p. ix), and aids communication about the objects of this quantification beyond the community itself. We can directly compare students on the measured trait, observe who has a higher level of intelligence or other cognitive skills, and based on that, make a decision who is accepted into college and who is not.

In sum, the quantification of individual differences has both a moral and a practical component: the moral component being the idea of equality of standards, and the practical component being the usefulness of quantifying information (rather than qualitative). The strong adherence to quantification in psychometrics also enables the psychometrician to claim a very specific ideal of objectivity to which we move on now.

6.5.2 Objectivity

Objectivity is perhaps the quintessential epistemic value in all sciences, whether physical (e.g., physics and chemistry) or social, like psychology or economics. But as recent work in history and philosophy of science shows, objectivity has several meanings and which of these meanings is prized is highly contingent on the history and the context (Douglas, 2004; Daston & Galison, 2007). Objectivity can mean being true to nature from God's eye point of view, or it can mean restricting personal preferences or desires of the inquirer, or may amount to just following transparent and mechanical procedures, and so on. There is no one true definition and exactly which of the many definitions a discipline settles on can reveal a great deal about its identity. Psychometrics is no exception, and the definition of objectivity it adopts is motivated by considerations that are very specific to psychometrics: it is considered risky to let any personal judgement taint any part of the testing process. The personal judgments by individuals can be informed by factors that have little to do with an examinee's ability, such as the examinee's gender, religion or social class, or simply by a general dislike for the examinee. Psychometricians would object to a situation in which teachers exclusively rely on their own judgement. In making a decision like whether a student has the talents and skills necessary for college, teachers should be informed by a proper assessment procedure that controls for possible biases.

Psychometricians, and specifically Rasch modelers, have operationalized objectivity in a particular way: judgments about a person's ability, and decisions based upon these judgments, should be invariant under a change of perspective. The requirement that conclusions should be invariant under exchanging observers, raters, interviewers, or items is made explicit in several formal definitions of psychometric models. For example, in the basic unidimensional latent variable model (Mellenbergh, 1994a), it does not matter (except for measurement error) which particular items are used to arrive at conclusions regarding the ordering of individuals (Grayson, 1988). Perhaps the best-known instance of this property is Rasch' concept of *specific objectivity* (1967; 1968; 1977): the requirement that the comparison of persons should not depend on which items are chosen and vice versa (see also Fischer, 1987). Similarly, Junker and Ellis (1997) conceptualized the idea of *exchangeability* of psychometric items. Along similar lines, decisions should not depend on any particular person (a *rater* in psychometric jargon; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Objectivity in psychometrics thus means the removal as far as possible of any leeway and any freedom from the users and the administrators of a test, as well as the notion of specific objectivity or measurement invariance, in which the estimation of a person's ability "must not depend on the items nor on the distribution and sampling of persons." (Kreiner, 2007, p. 281).

Objectivity often denotes a departure of values from science, or it is considered an epistemic value at the most. Though objectivity is certainly an epistemic value in psychometric research, objectivity is also a moral value, namely, a social commitment to fairness, which we will elaborate on below.

6.5.3 Fairness

Both values discussed above, the aim for objective and quantitative measurements, have a strong fairness component. Not only should measurement take place in an objective fashion to avoid any value judgments from people involved, but ultimately, the most qualified people should be awarded the highest test scores and be granted whatever was at stake (e.g., the job position or enrollment at a university). However, in the second half of the 20th century, there was a growing awareness that people with the highest ability do not automatically gain the highest test score. In 1979, in a case known as the *Larry P.* - case, the American court decided that standardized IQ tests were racially biased against African-American children, who were, due to their lower scores, placed in inferior education settings, which in turn only increased their isolation from a suitable high-quality learning environment. This and other rulings resulted in a complete ban against administering IQ-tests to African-American children for any special education purpose (Frisby & Henry, 2016). The case of Larry P. and other lawsuits have contributed to the awareness among psychometricians that tests, and even items, can discriminate against specific groups (e.g., gender, cultural or minority groups). Test or item fairness became an important topic in psychometric research from the 1970s onwards.

In psychometrics, unfairness is defined as the following: an item is unfair when respondents from different groups who have the same measured ability, show different probabilities for answering an item correctly. The main theory that was developed to identify unfair items carries the name Differential Item Functioning, or DIF. In case of DIF, measurement is confounded by other (irrelevant) variables and thus cannot be considered objective (Kreiner, 2007). When one or more items is affected by DIF and this has adverse consequences for a specific group, we can speak of item bias. In that case, the objectivity of the measure has been negatively affected. Psychometricians have developed various methods for identifying unfair items (e.g., Holland & Wainer, 1993) and have written software packages so researchers could test for DIF (e.g., Shealy, Stout, & Roussus, 1991). Fairness as conceptualized in psychometrics is thus a rather restricted sense of fairness, which only pertains to the fairness of specific items or tests, not to a wider social or political fairness (see for example Broome (1990) for a radically different conception of fairness).

In factor analysis, fairness was formalized as *measurement invariance* (Meredith, 1993), which means that, across groups, the instrument relates to the latent variable in the same way. Whereas DIF is usually item-specific (though there are instances where groups of items rather than just one item are considered), measurement invariance denotes the

complete absence of DIF in a test. Another example of fairness in psychometrics is test equating. Some tests are held repeatedly (like the SAT college admission tests in the US), and test scores of these tests should be comparable across different versions. As with DIF and measurement invariance, psychometricians would argue that test equating should be evaluated by using statistical methods, not by test makers or experts, and test equating is another main tradition in psychometric research. Commentators on psychometrics widely recognise these (and other) conceptualisations of fairness in testing practice (Dorans & Cook, 2016).

Though psychometricians have become more conscious of the possible unfairness in testing, this has not led to a rejection of testing as such. Actually, the opposite is the case: standardized testing, now including methods for calculating some fairness statistic, is still considered one of psychometrics' main contributions to society (Wijsen, forthcoming). A society that is not aware of the problems of fairness in testing, or worse, a society that does not endorse testing at all, is not aware of the danger of its alternatives. Most likely, testing has provided many people with opportunities that they otherwise would not have had, and standardized testing might have made society fairer than it would have been without.

6.5.4 Utility above Truth

In early psychometrics, psychometricians were seeking a new type of objective and quantitative psychological science, which they hoped would finally uncover the human mind. As we have shown, the aim for true knowledge about human psychology gradually faded over time. Psychometrics developed from a research area with an interest in the true measurement of psychological mechanisms into an area mostly involved with methods for data analysis and prediction. Utility, rather than truth, became psychometrics' new objective.

In contemporary psychometrics, it is considered valuable when research material finds its place in one or more applications. In fact, the possibility of application of psychometric research is perceived as more valuable than uncovering fundamental knowledge and building theory that perhaps has no immediate connection to such applications (Borsboom, 2006; Sijtsma, 2006). Rather than trying to explain human behavior, most contemporary psychometric research answers very practical questions. Is this test fair with respect to test-takers from different cultural backgrounds? What is the best method to estimate a person's ability? How do we control for a set of confounds? How can we compare the abilities of school children from different countries? Exceptions aside (e.g., explanatory item response theory; De Boeck & Wilson, 2004, and SEM), most psychometric models usually do not aim at explanation. Psychometrics is thus relatively practical with regards to the question it answers and does not often seek a fundamental understanding of a mechanism. Knowledge for the sake of knowledge, or truth for the

sake of truth, – a value in many of the sciences –, has gained little territory in contemporary psychometrics.

To illustrate the utility value in psychometrics, we can draw a comparison between psychometrics and an engineering approach – a comparison that has also been suggested for economics (Roth, 2002). Wilson and Gochyyev (2013) argue that, similar to engineering, psychometrics builds or constructs objects (the measures) and then develops models to analyze these measures, which they call 'reverse-engineering'. Psychometrics can thus be considered "a practical and engineering activity rather than as a basic science" (p. 3), in which basic science denotes seeking (fundamental) knowledge and understanding of mechanisms. Similar to engineering, one of the ultimate goals in psychometrics is that the research is used for *building* useful structures, such as new testing procedures, more advanced methods for data analysis in psychological science, or user-friendly software packages. And similar to engineering, these structures are considered more valuable than foundational psychological knowledge that has no direct link to output. The latter would be, in the case of engineering, the territory of the (theoretical) physicists, or in the case of psychometrics, the territory of the psychologists. This is not to say that there is no foundational research in psychometrics. *Psychometrika* especially publishes many articles on foundational problems in psychological measurement. However, these articles often pertain to mathematical or statistical proofs, rather than foundational *psychological* knowledge.

Besides an emphasis on building new structures that add onto the already elaborate testing technology, psychometrics has also become increasingly focused on prediction and data analysis than on actual measurement. As mentioned previously, psychometricians themselves do not necessarily believe in the existence of psychological attributes, nor in the direct connection between reality and psychometric models. But psychometric models prove to be useful for the analysis of all sorts of behavioral data, which are increasing in number and volume. So rather than being restricted to measurement as it was originally intended, many psychometric models can be used for a wide array of data analytic purposes. Since data analysis and prediction often have a practical motivation (e.g., data reduction or data visualization), the emphasis on data analysis and prediction in psychometrics aligns with the value of utility above truth.

The value of utility over truth is one of the factors that separate psychometrics from mathematical psychology, a different quantitative research area in psychology that was once affiliated with psychometrics, but which has from the 1960s onwards developed independently (Van Zandt & Townsend, 2010). Both mathematical psychology and psychometrics rely heavily on mathematics and are model-based, but the use and interpretation of the models differ strongly. Models in mathematical psychology try to uncover laws or mechanisms that describe human behavior, mostly in the areas of

perceptive and cognitive processes (Batchelder, 2010). So rather than a focus on individual differences and measurement, mathematical psychology aims at developing formal psychological theory about cognitive processes. The concept of truth is therefore more a priority for mathematical psychology, than it is for psychometrics.

In summary, we can extract two aspects to the utility rhetoric in psychometrics. First, the notion that practical impact matters – it is important that psychometric work finds its way to the practice of educational measurement or applied research –, and second, that contributing to theoretical – and with that we mean 'substantive' – questions matters less than finding an application.

6.6 Conclusion & Discussion

Early psychometrics was devoted to a new social order: based on the measurement of people's innate abilities, people would have jobs and hold positions in society which they were *meant* to hold. The political and social motives of psychometrics have gradually faded over time, and in contemporary psychometrics – through its devotion to technological solutions and its absence from public debates – the illusion is created that psychometrics is free from any such values. In this article, we have shown that this idea is incorrect: several values permeate psychometric research, among which a commitment to objectivity and quantity as means to a reliable science, a commitment to fairness and equality, and the value of utility above truth.

The value-free ideal is the idea that scientific research should be free from social or political values. For the sake of good research, moral values should not have a direct or indirect effect on any part of developing good knowledge (Reichenbach, 1951). Over the past decades, the value-free ideal has been disputed by several authors (Douglas, 2009; Betz, 2013; Elliott, 2017). Non-epistemic values are not only inescapable but can even be defensible and desirable, at least up to a certain degree (Anderson, 2004). Research should inform policy-making, hence, the incorporation of moral values is intrinsic to doing good research. And since psychometrics is so closely connected to important social infrastructures, it would be undesirable for psychometrics to be value-free, and for psychometrics to be unconscious of these values.

In fact, some psychometricians might entertain too little consciousness of values in psychometric research, and of the social and political reality that is intimately connected to their work. Psychometricians endorse a strong sense of fairness, but define it in terms of the technical solutions and tools they have developed to identify the items that are biased. And although these tools were a response to a specific type of unfairness in the testing industry, testing and psychometrics have contributed to other types of social unfairness or inequality as well. High test scores have become a more than desirable good in many societies – they are the gateway to receiving top-notch education – and several structures

are in place to encourage students to achieve those scores. The technology of testing has contributed to a social order that is not only determined by ability, but also by factors like Socioeconomic Status, geographic location, and access to education. The emphasis on a strictly meritocratic society – though seemingly committed to fair treatment and equality – has its own dangers, which have been pointed out by several authors (Gordon & Terrell, 1981; Lemann, 1999; Sokal, 1987; Stein, 2014). Whereas the debate on the merits and desirability of testing is ongoing, culminated now in the current debate on the suitability of established tests for college admission, psychometricians tend to remain largely absent from these debates, whereas their insights and expertise will most likely prove a valuable addition. It takes a political decision, a value in fact, to see yourself as a discipline that does not engage in such socio-political debates.

In making these claims, we have no intention of debunking psychometrics in any sense. Any science that is studying something valuable, is a science that is value-laden (almost by definition). However, we want to debunk perhaps a certain image of psychometrics that might be entertained by default, which is of it as a neutral and value-free science. A debunking that we hope is useful for psychometricians to have a more reflexive and realistic image of what they do, and finally in particular, to entertain the possibility that different values can motivate their work. They can entertain a much more social idea of fairness; they can ask the more foundational questions of whether meritocracy is the right way to organize a society. That psychometrics is not asking itself these questions is not inevitable. The professional responsibility of the psychometrician is possibly more significant than foreseen and has to encompass reflections on whether the values that currently animate psychometrics are the values that should animate future psychometrics.

Chapter 7

General Discussion

7.1 Aims and Main Findings

In this dissertation, I have taken psychometrics as the object of my investigation, and have used a number of approaches – that of the historian, that of the ethnographer, and that of the philosopher of science – to shed light on different aspects of this complex scientific discipline. The aim of this dissertation was to unpack psychometrics and draw different characterizations of the field, so that psychometrics becomes a more accessible discipline for researchers outside the psychometric community, such as historians, sociologists and philosophers of science. Even though, I used a number of methods and frameworks throughout this dissertation, there is a logic to it. What keeps this thesis from being overly eclectic is its object: the curious field of psychometrics is in all chapters the object of inquiry.

The academic genealogy in Chapter 2 traces back the lineages of presidents of the Psychometric Society, and shows that psychometrics is a structurally diverse field, with branches stemming from psychology and mathematics, with Wilhelm Wundt, William James, and Carl F. Gauss (among others) as ancestors. This chapter has an important preservationist purpose: it documents important connections among psychometrics and part of its historical development. The genealogies show that, over time, the influx of people with other backgrounds has increased, and that psychometrics is an inherently multidisciplinary and increasingly diverse field. The diversity of psychometrics also becomes clear in Chapter 3, which holds a qualitative analysis of interviews with Psychometric Society presidents. The interviews show that the diversity of psychometrics is reflected in how psychometricians practice their research, and how they perceive the challenges that are facing the field. Even though the presidents each practice psychometrics in a sometimes radically different way (some with the aim of improving psychological science, others by building software for users all around the world, and again others because of the joy they experience when solving a mathematical theorem), they all worry in some way about the future of psychometrics. For some, this is because psychometrics has trouble reconnecting with psychology, and for others, it is because psychometrics has trouble connecting with other contemporary disciplines, such as statistics and data science. Based on the interviews, we recommend psychometrics to become more explicit about its priorities and to engage more in public debate.

Chapter 4 gives an overview of three causal modeling traditions in psychometrics, each appropriate for a specific causal hypothesis: namely a reflective model, in which the latent variable is the underlying explanation for observed variables, the formative model, in which the latent variable is a an effect of the observed variables, and the network model, which models potentially causal relationships among a set of (often observed) variables. Chapter 4 addresses a number of controversies regarding causal modelling in psychometrics such as the distinction between data models and causal models, the problem

of generalization, and interpretational confounding. Even though the models mentioned in this chapter do not conclusively solve the incredibly complex problem of causality in psychology, they may contribute to a better understanding of psychological phenomena.

Chapter 5 argues in favor of a causal reading of the common factor model (a reflective model) rather than a statistical reading, if circumstances allow. A causal reading is not always possible – if there is no reason to believe that the common cause structure explains the data, a causal reading is simply not appropriate –, but when it is hypothesized that the underlying causal structure of the data is in fact that of the common cause, a causal interpretation of the CFM offers several benefits. A causal interpretation conforms with most research questions in which the goal is to *explain* the correlations between indicators rather than merely summarizing them; a causal interpretation of the factor model legitimizes the focus on *shared*, rather than unique variance of the indicators; and a causal interpretation of the factor model legitimizes the assumption of local independence. Important to note that Chapters 4 and 5 do not necessarily describe the practice of psychometric modeling, since it is rare that a psychometrician would speak of the causal relationships between latent variables and indicators. Instead, these chapters offer a causal *reading* or *interpretation* of psychometric models and highlight some possibilities for psychological theory building if one were to uphold a causal interpretation of these models.

Chapter 6 provides an analysis of values in contemporary psychometrics. Where early psychometrics was explicit about its social and political motives, contemporary psychometrics is not: it invites the image of being an abstract, technical, and socially uninvolved discipline. We show that this image is incorrect: contemporary psychometrics employs a number of values, e.g. the conceptualization of individual differences as quantitative (not qualitative) differences, the aim for objective measurement, the aim for fair measurement, and the preference for utility above truth. This chapter does not intend to conclude that psychometricians are in any way wrong in employing these values, in fact, a field that is doing something valuable is naturally committed to a set of values. However, we do wish to show that the image of a value-free psychometrics is mistaken. With this chapter, we hope to encourage psychometricians to become more aware and reflexive of the values they practice in their research.

7.2 Paradoxes of Psychometrics

Investigating psychometrics through an outsider's perspective has given me much food for thought over the past couple of years. What has occupied my thoughts the most throughout this project are a number of underlying tensions or even paradoxes in psychometrics that I'm still having trouble understanding. These paradoxes highlight some of the main issues and dilemmas in psychometric research. Though my understanding about

these paradoxes is not conclusive, they might help outsiders comprehend some of the complexities of the field, and they possibly form a fruitful basis for future research.

7.2.1 Where is the 'Psycho' in Psychometrics?

One of the topics that is possibly most frequently addressed in this dissertation (especially, in Chapters 2, 3, and 6), is the relationship between psychometrics and its neighboring fields, such as psychology, educational measurement, and statistics. These relationships are difficult to capture: in some ways these areas overlap, in some ways they are worlds apart. The paradox here comes down to the relative (in)dependence of psychometrics as a scientific discipline. Though psychometrics has grown out of psychology and its founders were mainly psychologists, contemporary psychometrics has its own agencies, its own journals, its own meetings, and at some universities, its own department. It has thus become a relatively independent field, and does not seem to rely much on psychology. And though the name 'psychometrics' implies a strong relationship with the measurement of psychological attributes, psychological measurement is certainly not its only goal. In some cases, and especially in contemporary psychometrics, the psychological content of psychometrics has faded to the background. For a field where 'psycho' features quite prominently in its name, its relationship with psycho-related issues is rather thin.

This dissertation gives a number of reasons for questioning the amount of actual 'psycho' content in psychometrics. Chapter 2 shows that the influx of people with a different educational (often mathematical) background, has increased over time. In Chapter 3, several presidents admitted to knowing very little of psychology or not being interested in it. If they collaborate with psychologists, it is often the psychologists who comes up with an interesting research question, and the psychometrician is there for methodological advice or any other supporting role. Psychometricians think radically different about the purpose of psychometric research and the role it has to play in psychological or other substantive research. Some consider psychometrics to have a supporting role: psychometrics should serve psychology. Others consider psychometrics a matter of statistics – a set of content-empty methods that can be applied to a variety of research problems –, and again another group has high hopes for a psychometrics that is integrated with psychological or educational research. And as we have seen in Chapter 6, psychometricians are more often led by the technological aspects or the utility of a problem, than by the substantive research question.

The question thus remains, how much psycho-content does psychometrics entail? I am not suggesting a name change here, or that psychometrics has to make some serious changes right away. What I am suggesting is that understanding the extent to which psychometrics still falls under psychological research and how the relationship with psychology has perhaps changed over time, helps putting psychometrics in some

perspective. If it is indeed the case that contemporary psychometric research does not really relate to psychological research anymore, the question rises what it is that these contemporary psychometricians actually do and what psychometricians pursue. Moreover, it problematizes how psychometrics should serve psychology if it has lost touch with psychology. Perhaps the conceptualization of contemporary psychometrics as the field that deals with the measurement of psychological attributes is simply no longer valid.

7.2.2 Psychometrics is Committed to Social Problems, but (Almost) Never Explicitly

As elaborated on in Chapter 6, early psychometrics was explicitly committed to solving social problems, and has played a substantial role in the rise of a national education, military testing during WWI, and in the selection of children for special education. The explicitness has faded, and psychometrics has become a mostly technical exercise that does not often specify a social problem it intends to solve. However, the controversy regarding psychological measurement and educational testing has anything but faded, and each year topics like the validity of test instruments or the undesirable consequences of mass testing are subject of national debates. In fact, as we speak, the University of California has decided to suspend the SAT and ACT tests as admission requirement. More often than not, psychometricians are surprisingly absent in this debate. Remarkably, discussions on psychometric practice or psychometrics' influence on our society are held in journals like *Assessment in Education: Principles, Policy & Practice* (e.g. Baird, Andrich, Hopfeneck, & Stobart, 2017) and *Educational Measurement: Issues and Practice*, not in *Psychometrika*. Psychometricians might not consider themselves public figures and find themselves in a more comfortable place when they are able to work on their models.

But this is exactly where the paradox come in: psychometricians might not always be explicit about their values or social commitment, but they certainly are implicitly committed to certain social ideas about for instance the role of assessment in education. Psychological and educational measurement have shaped the social structure of our society, and psychometricians actively contribute to this social structure. And as Chapter 3 shows, many find testing an important achievement in the history of psychometrics: they find a society that uses testing to select and evaluate people's skills a fairer and more equal place than a society that selects people on personal preference.

So, if psychometricians are at least implicitly socially committed, and if they know from research that for example the CITO test is a better alternative than having teachers (or even worse, parents) decide about the future school of a child, or that the SAT or ACT also protect university admissions from unfairness, then why are they so silent in these and other education-related debates? The response of the psychometrician might be that they are simply better at the research side of things, that they are not public figures, and that it is not their task to mingle in political or social discussions. But, in my view, as long as

psychometricians are in fact socially committed, and if they endorse a social structure in which psychological and educational measurement is key, these arguments do not really hold. I could perhaps even state that, because they have unique expertise that so few other people have, they in fact shóuld participate in these debates. Psychometricians know the biases in people's reasoning and how human factors can affect rational decision making. They could be a reasonable voice in a debate that is at least partly run by arguments that are not evidence-based.

Psychometrics as a scientific discipline could benefit from becoming an overall more reflective discipline. As we indicated in Chapter 6, a more reflective approach towards how psychometric research and society are intertwined, could help in giving psychometricians a better idea of goals that are worthwhile pursuing. The reflection on psychometrics' relationship with our social reality goes two ways. Psychometricians should engage more in the social and political debates to show how and why assessment is a better alternative (as opposed to for instance human judgement), but it also means that psychometricians should be more aware when something is not worth pursuing, perhaps because it has undesirable consequences for test takers and society at large.

7.2.3 Psychometrics as Engineering, not as Science

This is perhaps the most controversial of the paradoxes discussed here. Let me state first that it is by no means my intention to accuse psychometrics of being a pseudoscience, or anything of the kind. It is not my intention at all to accuse psychometricians of practicing their field incorrectly. In my research and interactions with psychometricians, I have experienced that psychometricians are incredibly devoted to their work, and especially to doing it well and thoroughly. They are certainly not the flashy types. What this paradox comes down to is the observation that an understanding of psychometric research as scientific practice is not most descriptive of psychometric practice. Understanding psychometrics as engineering however, rather than as a fundamental scientific discipline, seems more appropriate in some instances.

The quantitative nature of psychometrics and its reliance on statistics creates the impression that psychometrics is perhaps more reliable, more trustworthy, and more scientific in nature than some other disciplines in the social sciences. Psychometrics endorses several values that are classically associated with science, like objectivity, transparency, reliability, generalizability, and replicability. The objective nature of psychometrics gives reason to believe that psychometrics can indeed be characterized as a science. However, there are a number of reasons why understanding psychometrics as engineering can be more appropriate.

Psychometrics as engineering is a parallel that has been drawn before (Thissen, 2001; Wilson & Gochyyev, 2013), and this dissertation gives several indications that this

parallel is indeed relevant. As we have seen in Chapter 6, psychometricians value utility over truth. Psychometricians do not often speak of real-world mechanisms or discoveries; they do not speak of truth and of theory. Whether or not psychological attributes exist and how they relate to other attributes is often not addressed in psychometric research. This is illustrated in Chapter 3: most psychometricians I interviewed are not very active in constructing behavioral theory. They prefer to leave that up to the psychologist. Instead, many psychometricians like to *build*, as engineers do. They engage in building highly predictive models, user-friendly software packages, and advanced reliable measurement processes, such as Computer Adaptive Tests or item banks. In the eyes of the psychometrician, predictability and efficiency are often considered more important than notions like truth and theory.

Similar to engineering, psychometrics could not be as advanced as it is today without is reliance on many fundamental scientific principles. Psychometric models rely heavily on statistical or mathematical principles, similarly as to how building tunnels or buildings relies heavily on the principles of mechanics. Psychometrics is thus certainly a consequence or product of the fundamental work (which some psychometricians certainly engage in), but psychometric research as activity is in many cases more application-oriented (just like engineering is).

The paradox is that between the impression of psychometrics as among the most 'sciency' in the social sciences, and the conclusion that science is perhaps not the appropriate term to describe most of psychometrics' activities. Seeing psychometrics as an engineering activity sheds light on what it is that psychometricians actually do and why. For example, it explains why psychometricians do not often engage in theory construction: they want to see the results of their work in the shape of a 'building', something that can be used and that is an improvement on earlier measurement procedures, not something that may be theoretically sound but is untouchable.

7.3 Implications for Further Research

Though this dissertation addresses several important issues, dilemmas, and paradoxes in psychometrics, it has only slightly uncovered the lid, and it has probably raised more questions than it answers. Below, I will describe a number of starting points for research into the fundamentals and practice of psychometric science.

7.3.1 A Full Ethnography of Psychometrics

Chapter 3 explores how psychometricians perceive their own field. As such, it treated 'the psychometricians' as a group of people with its own specific practices, motivators, and culture, which resembles an ethnographic approach. During this project I certainly felt like I was an observer trying to gain access to a mostly unfamiliar world. Though Chapter

3 cannot be considered a full ethnographic study, it borrows elements of an ethnographic study (e.g., the interview as a form of data collection). However, psychometrics could benefit from a more in-depth ethnographic investigation, that makes use of a wider variety of ethnographic methods (not just interviews, but also observations).

Ethnography is "the study of social interactions, behaviors, and perceptions that occur within groups, teams, organizations, and communities" (p. 512, Reeves, Kuper, & Hodges, 2008), and is usually considered one of the main research methods of the (cultural) anthropologist. Through an ethnographic investigation, the aim is to gain understanding of the perspectives and practices of people within a specific culture. Ethnography has become a popular research method for Science and Technology Studies (STS) as well, usually referred to as *laboratory ethnography* (Hess, 2001; Stephens & Lewis, 2017). Laboratory ethnographies aim to show how scientific knowledge is produced or constructed in specific contexts (such as labs, field work, or political settings).

Psychometricians do not work in the stereotypical laboratory, as chemists or biologists sometimes do, but there are different locations in which psychometrics, in different shapes and forms, takes place. There is, as Stephens and Lewis (2017) call it, the 'dry' laboratory, which is where the mathematical or statistical work is conducted (the territory of the β-psychometrians). There are the locations or contexts where the processing of the psychometric work into assessment procedures or software programs takes place (sometimes done by the β-psychometricians, but this is also the arena of applied researchers, data science companies, or organizations like ETS). There are also the schools, clinics, and job offices where people take part in assessments, and lastly, there is the political arena where policymakers discuss how testing should operate in society.

This dissertation treats β-psychometricians in isolation, but it would actually be very relevant to explore what constitutes psychometric knowledge, how it is produced and how it traverses from context to context (from the lab to the applied researcher, to ETS, to the schools). We have seen in Chapter 6 that utility or the possibility for application is an important value for many psychometricians. However, as Chapter 3 tells us, psychometricians also worry about the gap they see between applied research or psychometric practice, and psychometric research. Somewhere along the line, somewhere between the development of the technical psychometric knowledge and its application, the transport of psychometric knowledge is blocked or hindered. An ethnographic exploration could contribute to understanding how psychometric knowledge is transformed into psychometric practice, and where and why this sometimes goes wrong.

A second reason for expanding the ethnographic investigation into psychometrics, is to improve our understanding of values in psychometrics. Chapter 6 discussed a number of values of the β-psychometricians. But, as mentioned earlier, psychometrics is a science that takes place in different shapes and forms, and in different contexts. The question that

arises here is whether for example applied researchers or people who work for testing agencies act on a different set of values than the β-psychometricians. What aspect of doing research – theory building, data analysis, engineering, solving technological problems, solving social problems – do they prioritize? A more in-depth analysis of the use of values in different aspects of psychometric research could also help explaining why the transfer of knowledge is sometimes so difficult.

It goes without saying that if one was do an ethnographic study of psychometrics, the focus on Psychometric Society presidents alone (the focus of Chapter 3) is simply too restricted. What people consider hurdles or challenges for psychometric research, or what values are worthwhile for psychometrics to pursue might significantly differ among researchers of different generations or of different backgrounds.

7.3.2 A Pragmatist Philosophy of Psychometrics

A realist philosophy of psychometrics makes sense in some ways (Borsboom, Mellenbergh, & Van Heerden, 2004; Hood, 2013; Maul, 2013; Van Bork, Wijsen, & Rhemtulla, 2017), and the common factor model specifically invites a realist interpretation. Chapters 4 and 5 both contribute to a realist or causal understanding of psychometric models. However, in practice, psychometric models are rarely given a realist interpretation. Psychometric models are often used for practical purposes, such as prediction or data summarization, rather than for theory building, which aligns with the utility above truth value we discussed in Chapter 6. The focus of psychometrics on practicality rather than truth or foundational knowledge invites a pragmatist philosophy of psychometrics – an approach that has so far been neglected in the literature.

After the American philosopher Charles S. Peirce first conceptualized an account of pragmatism in the 1870s, pragmatism was further developed by the grandfather of American psychology: William James. As we have seen in Chapter 2, some of the presidents of the Psychometric Society and many other prominent psychometricians are part of his academic lineage. Pragmatism for William James was a method for settling metaphysical disputes (James, 1907; 1909). According to his pragmatism, the main question we should ask ourselves in any dispute is "What difference would it practically make to anyone if this notion rather than that notion were true?" (p. 45, James, 1907). If there is no practical difference between both options, "all dispute is idle" (p. 45). If there is indeed a practical difference to be detected, it should be possible to show how that difference follows from one or the other side. Pragmatist scientists do not partake in a venture for truth, but focus on 'successful working' instead (Barnes-Holmes, 2000). The pragmatic method thus focuses more on the consequences of a decision or dispute, so on action, rather than on what precedes the results, or, the theoretical assumptions that have no direct effect.

Though there have been some pragmatist accounts or readings of psychological science and several calls for a more pragmatic social science (Baert, 2005; Fishman, 1999, Guyon, Kop, Juhel, & Falissard, 2018), a more in-depth pragmatist reading of psychological research has not received much attention. Due to James' centrality in the academic genealogy of Psychometric Society presidents, it is not unlikely that James' philosophy, directly or indirectly, also found traction in psychometrics. Figure 6 in Chapter 2 shows that James' lineage includes major psychologists like G. Stanley Hall, Lewis Terman, and Morris R. Cohen, who are the academic ancestors of psychometricians like Paul W. Holland, Lawrence Hubert, and Quinn McNemar. Two more clues for the possible influence of pragmatism on psychometrics are the importance of utility as a value as opposed to truth-seeking (see Chapter 6) and the tendency of most psychometricians not to engage in matters of psychological theory building (since that is part of the psychologist's tasks, see Chapter 3). A pragmatist interpretation of psychometrics might thus prove a fruitful investigation.

A pragmatist reading of psychometrics generates a number of interesting research questions. If psychometricians are not led by matters of truth, theory, or causality, what aspects are leading when psychometricians make decisions during the modeling process? Do psychometricians base their decisions or the evaluation of certain theories or models on 'successful working'? And what, according to psychometricians, is the definition of 'successful working'? On what basis do psychometricians choose a specific model or estimation procedure? And, if we take on a pragmatist reading of psychometric models, how then should we interpret psychometric models? The ways in which psychometric thinking resembles pragmatic thinking, what it means for a psychometric model to be understood as a pragmatic model, and the historical influence of pragmatism on psychometrics, are relevant research topics that probably offer more accurate descriptions and interpretations of psychometric research practice than Chapters 4 and 5 currently provide. The influence of pragmatist thinking on psychometric research has so far not been treated in the literature – exceptions being Maul, Irribarra & Wilson (2016) and Irribarra (2016) – but could prove a relevant avenue for new research that enlightens the practice of psychometric modeling.

7.3.3 The Performativity of Psychometric Concepts

Psychometric applications are incredibly influential, and how testing and the conceptualization of intelligence has shaped society in a number of ways has been the topic of a number of studies (Gordon & Terrell, 1981; Gould, 1981, Stein, 2014). The prevalence of testing and the importance that is given to intelligence have also influenced the psychology and behavior of people: scoring highly on tests is the gateway to success in our modern-day meritocratic society, and making sure those scores are achieved (in

terms of the devotion of students, the ambition of parents, and the rise of agencies that help one achieve those scores) are now part of our behavioral repertoire. In a way, the conceptualization of intelligence and the emphasis on testing can be considered 'looping effects' (Hacking, 1994). Because of the reification of these concepts and applications, human behavior and psychology have changed. Psychometric applications thus have performative power: the large-scale measurement of human abilities is not a neutral endeavor, it has made structural changes to social, political, and psychological factors.

The performative power of testing is not a new idea. However, the studies that have addressed these issues have often overlooked the influence of psychometric concepts on educational testing and society. In a response to Baird et al., (2017), Borsboom and Wijsen (2017) make a case for "an open and historically even-minded survey of psychometric modelling and educational testing" (p. 441), and opt for an investigation of the effects of two psychometric concepts: unidimensionality and measurement invariance or fairness. Unidimensionality pertains to the idea that a test is supposed to measure a unidimensional construct or latent variable, and measurement invariance (also described in Chapter 6) holds that across groups, the measurement instrument relates to the latent variable in the same way. Both psychometric concepts have found their way to the practice of educational measurement, and have become standards for test construction, even when they are both unlikely to hold. The assumption of unidimensionality expects a test to measure a single underlying construct which is the cause of the item responses. In reality, this only happens in a few occasions. Most often, a test measures a set of constructs, and in those cases, unidimensionality as a standard is inappropriate. Moreover, in a broader sense, the notion of unidimensionality might have contributed to the idea of reducing a person to a single number which can be objectively measured (this single number most often being intelligence). Though this last assumption is perhaps a stretch, the effect of unidimensionality on educational measurement is undeniable, and is deserving of further investigation.

According to Borsboom and Wijsen (2017), a similar thing happens with measurement invariance. In psychometrics, a test is measurement invariant when across groups, the instrument relates to the latent variable in the same way. If the underlying model is indeed that of the common cause, it makes sense that it is undesirable for different groups to relate differently to one item. These items are then removed from the test. However, it is unlikely that what underlies item responses is indeed that one single factor. What happens in practice is that items that behave differently are still removed, but it could actually be informative that different groups score differently on the same test. Hence, measurement invariance, and with that notions related to psychometric fairness such as differential item functioning, have an impact on test construction and perhaps more importantly, the way we understand what it means for a test to be fair.

Besides the effect of psychometric concepts like unidimensionality and measurement invariance, Borsboom and Wijsen (2017) also suggests a postmodern Foucauldian-like analysis of what we call *psychometric power*. Such a Foucauldian analysis would question how psychometrics exerts power over society. Psychometric applications regulate many parts of our lives (who goes to which university, who receives a diagnosis for depression, who is incarcerated, who gets the job). Psychometrics has thus contributed to an *aristocratic meritocracy*, and legitimizes this ideology through a huge body of highly technical and reliable research. All in all, the influential position of psychometrics in society, and the effect of some of its concepts on the practice of educational measurement, call for a more critical reflection on the ideology of psychometric research and how it exerts its power.

7.3.4 Discipline Formation of Psychometrics

This dissertation provides food for thought for sociologists of science with a specific interest in the development of the social sciences or psychology. Sociology of science is concerned "with the social conditions and effects of science, and with the social structures and processes of scientific activity" (p. 203, Ben-David & Sullivan, 1975). Sociologists of science investigate matters like the organization of scientific activity and the emergence of new disciplines. Over time, the sociology of science has provided many insights on the process of discipline formation (e.g. Fuchs, 1992; Stichweh, 1992) and discipline specialization (Ben-David & Collins, 1966; Price, 1963; Wray, 2005). The explanations for scientific specialization could be roughly divided into social factors (e.g. crowding in a field or the possibility of new job opportunities elsewhere), and conceptual factors (e.g. finding uncovered ground or a major discovery).

This dissertation gives a number of stepping stones for an investigation into the discipline formation of psychometrics, and hints at why such an investigation might be worthwhile. This largely has to do with the formation of psychometrics as a discipline apart from psychology and education. For example, in Chapter 2, we show that psychometrics is inherently a multidisciplinary area, and suggest that psychometrics' multidisciplinary character is at least partly responsible for the tension between substantive and technical orientations. This tension shows in the split of the Psychometric Society with two other organizations: The Society for Mathematical Psychology and the Society for Experimental Psychology, both of which have a stronger substantive orientation than the Psychometric Society. The question that rises here is whether both split-offs and the gradual detachment from psychology might have had a significant effect on the identity of psychometric research, and also on the position of psychometrics among its neighbors.

Besides the tension between psychometrics and psychology, Chapter 3 also identifies a number of other possible pitfalls, future challenges, and approaches in contemporary psychometrics, such as "an influx of researchers with a different educational background,

less familiarity with traditional research topics, an increasing detachment between people with substantive interests and people who are more technology and statistics oriented, a more diverse playground of research topics" (p. 75-76). Throughout the 20th century, psychometrics has thus made a number of maneuvers that have socially and conceptually changed the structure of the discipline, which provides a highly interesting and relevant avenue for a sociological analysis.

Can we indeed speak of an independent psychometric discipline? When did it become its own discipline (was that already in the early 20th century or is that a more recent development?) and what are the conceptual, cognitive, and social changes that led to psychometrics' independence? Moreover, can we identify a pattern of disciplinary development or specialization that is specific for psychometrics, and is this a pattern that we can transpose to other disciplines in a similar stage of development? Can we find a general tendency for research domains that started out as being part of a larger general discipline to develop as a largely separate discipline – with its own journals, lab groups, institutions, concepts – and lose touch with their original home base? A sociological analysis of the discipline formation of psychometrics will increase our understanding of how psychometrics has developed over time, and also contributes to our understanding of discipline specialization.

7.4 Conclusion

All in all, this dissertation shows that contemporary psychometrics is a structurally multidisciplinary research domain, which consists of a variety of different approaches and purposes, and which experiences a number of inner tensions. In psychometrics, we find an intricate balance between contributing to substantive psychological or educational research on the one hand, and being a mostly technical discipline on the other hand. Though psychometric modeling can certainly contribute to substantive research, and some models are indeed useful for explaining psychological phenomena, few psychometricians consider psychometrics suitable for that purpose. Instead, psychometrics can be seen as a relatively pragmatic discipline that values real-world applications and research that is technologically advanced.

I hope that this dissertation can contribute to a better understanding of psychometrics as a scientific discipline, that it has unlocked some of the complexities of psychometric research, and that it has made psychometrics an altogether more accessible discipline. Ultimately, these studies show that even the small discipline of psychometrics harbors many values, purposes, and concepts that are worthy of further investigation, and I hope that they encourage a more thorough reflection on what it actually is that psychometricians (ought to) do.

References

Abrams, L. (2010). *Oral History Theory*. New York: Routledge.

Anderson, E. (2004). Uses of value judgments in science: A general argument, with lessons from a case study of feminist research on divorce. *Hypatia, 19,* 1 - 24.

Arntzenius, F. (1993). The common cause principle. *Philosophy of Science Association 1992*, *2*, 227 -237.

Asmundson, G. J.G., Frombach, I., McQuaid, J., Pedrelli, P., Lenox, R., & Stein, M. B. (2000). Dimensionality of posttraumatic stress symptoms: a confirmatory factor analysis of DSM-IV symptom clusters and other symptom models. *Behaviour Research and Therapy, 38*, 203 - 214.

Baer, M. A., Jewell, M., & Sigelman, L. (Eds.). (1991). *Political Science in America: Oral Histories of a Discipline*. Lexington: University Press of Kentucky.

Baert, P. (2005). Towards a pragmatist-inspired philosophy of social science. *Acta Sociologica, 48,* 191 – 203.

Bainter, S. A., & Bollen, K. A. (2014). Interpretational confounding or confounded interpretations of causal indicators?. *Measurement: Interdisciplinary Research & Perspectives, 12,* 125 - 140.

Baird, J., Andrich, D., Hopfeneck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice, 24,* 317 – 350.

Barberousse, A. & Ludwig, P. (2009). Models as fictions. In M. Suarez (Ed.), *Fictions in science: Philosophical essays on modeling and idealization.* New York: Routledge.

Barnes, B. (1977). *Interests and the growth of knowledge.* London: Routledge Kegan Paul.

Barnes-Holmes, D. (2000). Behavioral pragmatism: No place for reality and truth. *The Behavior Analyst, 23,* 191 – 202.

Batchelder, W. H. (2010). Mathematical psychology. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 759 - 765.

Baumgartner, H., & Homburg, C., (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13,* 139 – 161.

Ben-David, J., & Collins, R. (1966). Social factors in the origins of a new science: The case of psychology. *American Sociological Review, 31,* 451 – 465.

Ben-David, J., & Sullivan, T. A. (1975). Sociology of science. *Annual Review of Sociology, 1,* 203 – 222.

Bennett, A. F., & Lowe, C. (2005). The academic genealogy of George A. Bartholomew. *Integrative and Comparative Biology, 45,* 231 – 233.

Bennett, R. E., & Von Davier, M. (Eds.). (2017). *Advancing human assessment: The methodological, psychological and policy contributions of ETS.* New York: Springer.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology, 31,* 419 - 456.

Bentler, P. M. (1995). *EQS structural equation modeling manual*. Encino, CA: Multivariate Software.

Berliner, D. C. (1993). The 100-year journey of educational psychology: From interest to disdain to respect for practice. In T. K. Faigin & G. R. Vandenbos (Eds.), *Exploring applied psychology: Origins and critical analysis* (pp. 39 - 78). Washington, DC: American Psychological Association.

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science, 3,* 207 – 220.

Blalock, H. M. (1985). *Causal models in the social sciences.* New York: Aldine Publishing Company.

Block, N. J. (1974). Fictionalism, functionalism and factor analysis. In R. Cohen (Ed.), *Boston Studies* (pp. 127 - 141). Dordrecht: Reidel.

Blumenthal, A. L. (1977). Wilhelm Wundt and early American psychology: A clash of two cultures. *Annals of the New York Academy of Sciences, 291,* 13 – 20.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53,* 605 - 634.

Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly, 35,* 359 - 372.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*, 265 - 284.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305 - 314.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation modeling. In S. L. Morgen (Ed.), *Handbook of causal analysis for social research* (pp. 301 – 328). Dordrecht, Netherlands: Springer.

Bollen, K. A., & Ting, K. F. (1993). Confirmatory tetrad analysis. *Sociological Methodology*, *23*, 147 - 175.

Bollen, K. A., & Ting, K. F. (2000). A tetrad test for causal indicators. P*sychological Methods*, 5, 3 - 22.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.).* Mahwah, NJ: Lawrence Erlbaum.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425 – 440.

Borsboom, D. (2008). Latent variable theory. *Measurement*, *6*, 25 - 53.

Borsboom, D. (2015). What is causal about individual differences?: A comment on Weinberger. *Theory & Psychology, 25,* 362 - 368.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*, 91 - 121.

Borsboom, D., & Dolan, C. V. (2006). Why *g* is not an adaptation: A comment on Kanazawa (2004). *Psychological Review, 113*, 433 - 437.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203 - 219.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004.) The concept of validity. *Psychological Review, 111*, 1061 - 1071.

Borsboom, D., & Wijsen, L. D. (2017). Psychology's atomic bomb. *Assessment in Education: Principles, Policy & Practice, 24,* 440 – 446.

Bos, E. H., & Jonge, P. De. (2014). "Critical slowing down in depression" is a great idea that still needs empirical proof. *Proceedings of the National Academy of Sciences, 111,* E878.

Bos, E. H., & Wanders, R. B. K. (2016). Group-level symptom networks in depression. *JAMA Psychiatry*, 73, 411.

Bowling, N. A., & Hammond, G. D. (2008). A meta-analytic examination of the construct validity of the Michigan Organizational Assessment Questionnaire Job Satisfaction Subscale. *Journal of Vocational Behavior, 73,* 63 – 77.

Broome, J. (1990). Fairness. *Proceedings of the Aristotelian Society, 91,* 87 - 101.

Burt, C. (1909). Experimental tests of general intelligence. *British Journal of Psychology, 3,* 94 - 177.

Burt, C. (1955). The evidence for the concept of intelligence. *British Journal of Educational Psychology, 25,* 158 - 177.

Burt, C. (1957). Inheritance of mental ability. *Nature, 179,* 1325 - 1327.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Buss, A. R. (1976). Galton and the birth of individual differences and eugenics: Social, political and economic forces. *Journal of the History of Behavioral Sciences, 12,* 47 - 58.

Buswell, G. T. (1947). Charles Hubbard Judd: 1873 – 1946. *The American Journal of Psychology, 60,* 135 – 137.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Carroll, J. D., & Arabie, P. (2013). In memoriam Joseph B. Kruskal 1928 – 2010. *Psychometrika, 78*, 237 – 239.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., . . . Mott, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*, 119 - 137.

Cattell, J. M. (1890). V.- Mental tests and measurement. *Mind, 59,* 373 - 381.

Cattell, R. B. (1990). The birth of the Society of Multivariate Experimental Psychology. *Journal of the History of the Behavioral Sciences, 26,* 48 – 57.

Chitty, C. (2007). *Eugenics, race, and intelligence in education.* London: Continuum.

Compas, B. E., Boyer, M. C., Stanger, C., Colletti, R. B., Thomse, A. H., Dufton, L. M., & Cole, D. A. (2006). Latent variable analysis of coping, anxiety/depression, and somatic symptoms in adolescents with chronic pain. *Journal of Consulting and Clinical Psychology, 74,* 1132 – 1142.

Coombs, C. H. (1952). *A theory of psychological scaling (Vol.34).* Ann Arbor: Engineering Research Institute, University of Michigan.

Coombs, C. H. (1964). *A Theory of Data*. Oxford, England: Wiley.

Cramer, A. O., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., … Borsboom, D. (2012). Measurable like temperature or mereological like flocking? On the nature of personality traits. *European Journal of Personality, 26,* 451 – 459.

Cramer, A. O., Waldorp, L. J., Van der Maas, H. L., & Borsboom, D. (2010). Complex realities require complex theories: Refining and extending the network approach to mental disorders. *Behavioral and Brain Sciences, 33*, 178 - 193.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297 – 334.

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychology tests. *Psychological Bulletin, 52,* 281 – 302.

Cronbach, L. J., Rajaratnam, N. & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137 – 163.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported Aptitude x Treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 57,* 717 - 724.

Danziger, K. (1994). *Constructing the subject: Historical origins of psychological research*. Cambridge: Cambridge University Press.

Danziger, K. (1997). *Naming the mind: How psychology found its language.* Thousand Oaks, CA: Sage.

Daston, L., & Galison, P. (2007). *Objectivity.* MIT Press: New York.

David, S. V., & Hayden, B. Y. (2012). Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PLoS ONE, 7,* doi:10.1371/journal.pone.0046608.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Leeuw, J. (1984). The Gifi system of nonlinear multivariate analysis. In E. Diday (Ed.), *Data analysis and informatics III (*pp. 415 – 424). Amsterdam: North Holland Publishing Company.

Dehue, T. (1995). *Changing the rules: Psychology in the Netherlands, 1900 – 1985.* Cambridge: Cambridge University Press.

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, *61*, 1203 - 1218.

Doel, R. E. (2003). Oral history of American science: A forty-year review. *History of Science, 41,* 349 – 378.

Dorans, N. J. (2011). Holland's advice for the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans & S. Sinharay (Eds.), *Looking Back: Proceedings of a conference in honor of Paul W. Holland* (pp. 259 – 272). New York: Springer.

Dorans, N. J., & Cook, L. L. (Eds.) (2016). *Fairness in educational assessment and measurement*. New York: Routledge.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science, 67,* 559 – 579.

Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese, 138*, 453 - 473.

Douglas, H. (2009). *Science, policy, and the value-free ideal.* University of Pittsburgh Press.

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2015). Political diversity will improve social psychological science. *Behavioral and brain sciences, 38,* E130. doi:10.1017/S0140525X14000430

Dubois, P. H. (1970). *A history of psychological testing*. Boston, MA: Allyn and Bacon.

Dworkin, R. (1981). What is equality? Part 2: Equality of resources. *Philosophy & Public Affairs, 10,* 283 - 345.

Edwards, A. L. (1954). *Manual for the Edwards Personality Preference Schedule.* New York: Psychological Corporation.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155 – 174.

Elliott, K. C. (2017). *A tapestry of values: An introduction to values in science.* Oxford University Press.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Erlbaum.

Epskamp, S. (2020). psychonetrics: Structural equation modeling and confirmatory network analysis. R-package version 0.7.2

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. In Irwing P., Hughes D., & Booth T. (Eds.), *The Wiley Handbook of Psychometric Testing,*

2 Volume Set: A Multidisciplinary Reference on Survey, Scale and Test Development. New York, NY: Wiley.

Epskamp, S., & Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82,* 904 – 927.

Evans, B., & Waites, B. (1981). *IQ and mental testing: An unnatural science and its social history.* London: Macmillan Press.

Fienberg, S. E., Hoaglin, D. C. & Tanur J. M. (2013). *Frederick Mosteller 1916 – 2006: A biographical memoir*. Washington, DC: National Academy of Sciences.

Fine, A. (1993). Fictionalism. *Midwest Studies in Philosophy, 18,* 1-18.

Fischer, G. A. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. *Psychometrika, 52,* 565 - 587.

Fishman, D. (1999). *The case for pragmatic psychology.* New York: University Press.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327 – 358.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286 – 299.

Frigg, R., & Hartman, S. (2005). Models in science. In S. Sarkar & J. Pfeifer (Eds.), *The philosophy of science: An encyclopedia.* (pp. 740 - 749). New York: Routledge.

Frisby, C. L., & Henry, B. (2016). Science, politics and best practices: 35 years after Larry P.. *Contemporary School Psychology, 20,* 46 - 62.

Fuchs, S. (1994). *The professional quest for truth: A social theory of science and knowledge.* State University of New York Press.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences.* London: Macmillan.

Galton, F. (1879). Psychometric experiments. *Brain, 2,* 149 - 162.

Gargiulo, F., Caen, A., Lambiotte, R., & Carletti, T. (2016). The classical origin of modern mathematics. *EPJ Data Science, 5,* doi:10.1140/epjds/s13688-016-0088-y

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology, 26,* 309 - 320.

Goertzen, J. R. (2008). On the possibility of unification: The reality and nature of the crisis in psychology. *Theory & Psychology, 18,* 829 – 852.

Gordon, E. M., & Terrell, M. D. (1981). The changed social context of testing. *American Psychologist, 36*, 1167 - 1171.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383 - 392.

Green, B. F. (1980). *Ledyard R. Tucker's affair with psychometrics: The first 45 years*. Princeton, NJ: ETS. Green, B. F. (1986). Models, computers and policies: Fifty years of Psychometrika. *Psychometrika, 51*, 65 - 68.

Green, B. F. (1999). Obituary: Warren S. Torgerson, 1924 – 1999. *Psychometrika, 64*, 3 – 4.

Greenwood, J. D. (2015). *A conceptual history of psychology: Exploring the tangled web*. Cambridge University Press.

Groenen, P. J., & Van der Ark, A. J. (2006). Visions of 70 years of psychometrics: The past, present and future. *Statistica Neerlandica, 60,* 135 – 144.

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin, 53,* 267 – 293.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139 – 150.

Guttman, L. (1953). Image theory for the structure of quantitative variates. *Psychometrika, 18,* 277 – 296.

Guttman, L. (1956). "Best possible" systematic estimates of communalities. *Psychometrika, 21,* 273 – 285.

Guttman, L. (1960). The matrices of linear least-squares image analysis. *British Journal of Statistical Psychology, 13,* 109 – 118.

Guyon, H., Kop, J. L., Juhel, J., Falissard, B. (2018). Measurement, ontology and epistemology: Psychology needs pragmatism-realism. *Theory & Psychology, 28,* 149 – 171.

Hacking, I. (1994). The looping effects of human kinds. In D. Sperber, D. Premack & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary approach.* Oxford, England: Clarendon Press.

Haggbloom, S. J., Warnick, R., Warnick, J. E., Jones, V.K., Yarbrough, G. L., Russell, T. M., … & Monte, E. (2002). The 100 most eminent psychologists of the 20th century. *Review of General Psychology, 6,* 139 – 152.

Haig, B. D. (2005a). An abductive theory of scientific method. *Psychological methods, 10,* 371 – 388.

Haig, B. D. (2005b). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research, 40*, 303 - 329.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge, MA: MIT Press.

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43 - 61). New York, NY: Guilford Press.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*, 102 – 116.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Press.

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2020). *Modeling psychopathology: From data models to formal theories.* PxyArXiv. https://doi.org/10.31234/osf.io/jgm7f

Heath, R. A. (2000). *Nonlinear dynamics: Techniques and applications in psychology.* Mahwah, NJ: Lawrence Erlbaum.

Heiser, W. J. (2008). Psychometric Roots of Multidimensional Data Analysis in the Netherlands: From Gerard Heymans to John van de Geer. *Electronic Journal for History of Probability and Statistics, 4,* 1 - 25.

Heiser, W. J. (2012). In memoriam, J. Douglas Carroll 1939 – 2011. *Psychometrika, 78*, 5 – 13.

Heiser, W. J. (2017). *Early psychometric contributions to Gaussian Graphical Modelling: A tribute to Louis Guttman.* Paper presented at the International Meeting of the Psychometric Society, Zurich.

Heiser, W. J., & Hubert, L. (2016). A creation narrative for the Psychometric Society and *Psychometrika*: In the beginning there was Paul Horst. *Psychometrika, 81,* 1172 – 1176.

Heiser, W., Hubert, L., Kiers, H., Köhn, F.-F., Lewis, C., Meulman, J., … & Takane, Y. (2016). Commentaries on the ten most highly cited Psychometrika articles from 1936 to the present. *Psychometrika, 81,* 1177 – 1211.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Hess, D. (2001). Ethnography and the development of Science and Technology Studies. In: Atkinson, P., Coffey, A., Delamont, S., Loftland, J., and Loftland, L., (Eds.), *Handbook of Ethnography* (pp. 234 – 245). London: Sage.

Hilgard, E. R. (1956). Robert Mearns Yerkes, 1867 – 1956, A biographical memoir. *Biographical Memoirs, 36,* 385 – 425.

Hilgard, E. R. (1987). *Psychology in America: A historical survey*. Orlando, FL: Harcourt, Brace, Jovanovich.

Hirshman, B. R., Tang, J. A., Jones, L. A., Proudfoot, J. A., Carley, K. M., Marshall, L., … & Chen, C. C. (2016). Impact of medical academic genealogy on publication patterns: An analysis of theliterature for surgical resection in brain tumor patients. *Annals of Neurology, 79,* 169 – 177.

Ho, Y.S., & Hartley, J. (2016). Classic articles in Psychology in the *Science Citation Index Expanded*: A bibliometric analysis. *British Journal of Psychology, 107*, 768 – 780.

Holland, P. W., & Wainer, H. (Eds.). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Holzinger, K. J., & Harman, H. H. (1941). *Factor analysis. A synthesis of factorial methods.* Chicago: University Press.

Hood, S. B. (2008). Latent variable realism in psychometrics. Unpublished doctoral dissertation, Indiana University, Bloomington, Indiana.

Hood, S. B. (2013). Psychological measurement and methodological realism. *Erkenntnis, 78,* 739 – 761.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12*, 205 - 218.

Hubert, L. (2012). In memoriam: Phipps Arabie 1948 - 2011. *Journal of Classification, 29,* 260 – 261.

Hubert, L., Arabie, P., & Meulman, J. (2001). *Combinatorial Data Aanalysis: Optimization by Dynamic Programming*. Philadelphia, PA: SIAM.

Hull, C. L. (1944). Joseph Jastrow: 1863 – 1944. *The American Journal of Psychology, 57*, 581 – 585.

Inquirium, L. L. C. (2013). *Inqscribe: Digital media transcription software.* Retrieved from http://www.inqscribe.com/ on August 30th, 2016.

Jackson, J. P., & Weidman, N. M. (2004). *Race, racism, and science: Social impact and interaction.* Santa Barbara, CA*:* ABC-CLIO.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (Vol. 112).* New York, NY: Springer.

James, W. (1907). *Pragmatism: A new name for some old ways of thinking*. Cambridge, MA: Harvard University Press.

James, W. (1909). *Pragmatism and the meaning of truth*. Cambridge, MA: Harvard University Press.

Jensen, A. R. (1999). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 1 - 27). New York, NJ: Elsevier.

Jonas, K. G., & Markon, K. E. (2016). A descriptivist approach to trait conceptualization and inference. *Psychological Review*, *123*, 90 - 96.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*, 239 – 251.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In Goldberger, A.S. & Duncan, O. D. (Eds.), *Structural equation models in the social scienc*es (pp. 83 – 112). New York: Academic Press.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631 - 639.

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics, 25*, 1327 - 1343.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 24,* 401 – 415.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement, 34,* 111 – 117.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319 - 342.

Kealy, W. A., & Mullen, C. A. (1996). Re-thinking mentoring relationships. Paper presented at the annual meeting of the American Educational Research Association, New York. (Microfiche No. ED 394420)

Kelley, E. A., & Sussman, R. W. (2007). An academic genealogy on the history of American field primatologists. *American Journal of Physical Anthropology, 132,* 406 – 425.

Kelley, T. L. (1923). *Statistical method.* New York: Macmillan.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmillan.

Kelley, T. L. (1947). *Fundamentals of Statistics*. Cambridge, MA: Harvard University Press.

Kievit, R. A., Romeijn, J. W., Waldorp, L. J., Wicherts, J. M., Scholte, H. S., & Borsboom, D. (2011). Mind the gap: A psychometric approach to the reduction problem. *Psychological Inquiry*, *22*, 1 - 21.

Kinderman, R. P., & Snell, J. L. (1980). On the relation between Markov random fields and social networks. *Journal of Mathematical Sociology, 7,* 1 – 13.

Kreiner, S. (2007). Validity and objectivity: Reflections on the role and nature of Rasch models. *Nordic Psychology, 59,* 268 - 298.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, *56*, 921 - 926.

Kruskal, J., & Wish, M. (1978). *Multidimensional scaling.* Newbury Park, CA: Sage.

Latour, B. (1987). *Science in Action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

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

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science, Vol. 3* (pp. 476 – 543). New York: McGraw Hill.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* New York: Houghton Mifflin.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy.* New York: Farrar, Strauss, and Giroux.

Levelt, W. J. M., Van de Geer, J. P., & Plomp, R. (1966). Triadic comparisons of musical intervals. *The British Journal of Mathematical and Statistical Psychology, 19,* 163 – 179.

Lindley, D. V. (1987). Melvin R. Novick: His work in Bayesian statistics. *Journal of Educational Statistics, 12,* 21 – 26.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental testing.* Reading, MA: Addison-Wesley.

Longino, H. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, Science and the Philosophy of Science* (pp. 39 - 58). Dordrecht: Kluwer.

Lorge, I., & Thorndike. R. L. (1957). *The Lorge-Thorndike Intelligence Tests.* New York: Houghton Mifflin.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*, 185 - 199.

Maraun, M. D. (1996). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research, 31,* 603 - 616.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of validity theory: Measurement, causation and meaning.* New York, NY: Taylor & Francis.

Marsh, H. W., Ludtke, O., Muthèn, B. O., Asparouhov, T., Morin, A. J. S., Trautwein, U. (2010). A new look at the Big Five structure through exploratory structural equation modeling. *Psychological Assessment, 22*, 471 – 491.

Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology, 23,* 752 – 769.

Maul, A., Irribarra, D. T., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79,* 311 – 320.

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81 – 90.

McCrae, R. R., & Costa, P. T., Jr. (2008). Empirical and theoretical status of the five-factor model of personality traits. In G. Boyle, G. Matthews & D. Saklofske (Eds.), *The Sage handbook of personality theory and assessment* (Vol. 1, pp. 273 - 294). Los Angeles, CA: Sage.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics, 30*, 1225 - 1267.

McNemar, Q. (1949). *Psychological Statistics.* New York: Wiley.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127 – 143.

Mellenbergh, G. J. (1994a). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29,* 223 - 236.

Mellenbergh, G. J. (1994b). Generalized linear item response theory. *Psychological Bulletin, 115,* 300 - 307.

Meredith, W. (1993). Measurement invariance, factor analysis and factor invariance. *Psychometrika, 58,* 525 – 543.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955 - 966.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5 - 11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13 - 103). New York: Macmillan.

Messick, S. (1998). Harold Oliver Gulliksen (1903-1996): Obituary. *American Psychologist*, *53*, 564 – 565.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* Cambridge University Press.

Michell, J. (2003). The quantitative imperative: Positivism, naive realism and the place of qualitative methods in psychology. *Theory & Psychology, 13,* 5 - 31.

Miles, W.R. (1956). *Raymond Dodge, 1871-1942: A Biographical Memoir*. National Academy of Sciences, Washington DC.

Minton, H. (1987). Lewis M. Terman and mental testing: in search of a democratic ideal. In M. Sokal (Ed.), *Psychological testing in American society: 1890 - 1930* (pp. 95 - 113). New Brunswick: Rutgers University Press.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2,* 201 - 218.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of inter-individual and intra-individual variability: A theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development.* Dordrecht, the Netherlands: Kluwer.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14,* 5 - 13.

Moulines, C. O. (2016). Patrick Suppes: A Profile. *Journal for General Philosophy of Science, 47*, 1 – 10.

Moneta, A., & Russo, F. (2014). Causal models and evidential pluralism in econometrics. *Journal of Economic Methodology, 21*, 54 - 76.

Mulkay, M. J. (1976). Norms and ideology of science. *Social Science Information, 15,* 637 – 656.

Mulaik, S. A. (1972). *Foundations of factor analysis.* New York: McGraw-Hill.

Mulaik, S. A. (1996). On Maraun's deconstructing of factor indeterminancy with constructed factors. *Multivariate Behavioral Research, 31,* 579 - 592.

Musek, J. (2007). A general factor of personality: Evidence for the big one in the five-factor model. *Journal of Research in Personality, 41*, 1213 - 1233.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment.* London: Sage.

Nafstad, H. E., & Blakar, R. M. (2012). Ideology and social psychology. *Social and Personality Compass, 6,* 282 – 294.

Norrgard, K. (2008). Human testing, the eugenics movement, and IRBs. *Nature Education, 1,* 170.

Novick, M. R., Hamer, R. M., & Chen, J. J. (1979). The Computer-Assisted Data Analysis (CADA) Monitor (1978). *The American Statistician, 33,* 219 – 220.

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

Price, D. J. De Solla. (1963). *Little science, big science.* New York: Columbia University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainments tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1967). An informal report on a theory of objectivity in comparisons. In L. J. Th. van der kamp & C. A. J. Vlek (Eds.), *Measurement theory.* Proceedings of the NUFFIC international summer session in science in "Het Oude Hof", The Hague, July 14 - 28, 1966. Leiden: University of Leiden.

Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. *Report from European Meeting on Statistics, Econometrics and Management Science, Amsterdam.*

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14,* 58 - 94.

Redding, R. E. (2001). Sociopolitical diversity in psychology: The case for pluralism. *American Psychologist, 56,* 205 - 215.

Reed, J. (1987). Robert M. Yerkes and the mental testing movement. In M. M. Sokal (Ed.), *Psychological testing and American society 1890 - 1930* (pp. 75 - 94). New Brunswick: Rutgers University Press.

Reeves, S., Kuper, A., & Hodges, B. D. (2008). Qualitative research methodologies: Ethnography. *British Medical Journal, 337,* 512 – 514.

Reichenbach, H. (1951). *The rise of scientific philosophy*. Los Angeles: University of California Press.

Reichenbach, H. (1956). *The direction of time*. Los Angeles: University of California Press.

Richards, G. (1997). "*Race", racism and psychology: Towards a reflexive history.* London: Routledge.

Robertson, J. M. (1994). Tracing ideological perspectives through 100 years of an academic genealogy. *Psychological Reports, 75,* 859 – 879. Rooney, P. (1992). On values in science: Is the epistemic/non-epistemic distinction useful? *Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1,* 13 - 22.

Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica, 70,* 1341 - 1378.

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology, 31*, 43 - 53.

Schumacker, R. E., & Marcoulides, G. A. (Eds.). (1998). *Interaction and nonlinear effects in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum.

Sears, R. R. (1961). Carl Iver Hovland: 1912-1961. *The American Journal of Psychology, 74*, 637 – 639.

Shalizi, C. (2013). *Advanced data analysis from an elementary point of view.* Cambridge University Press.

Shealy, R., Stout, W. F., & Roussos, L. (1991). *SIBTEST user manual* [Computer program manual]. Champaign: University of Illinois, Department of Statistics.

Shrader, R. R. (1994). Edward E. (Ted) Cureton (1902-1992). *American Psychologist, 49,* 350.

Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science? *Psychometrika, 71,* 451 – 455.

Sijtsma, K. (2009). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika, 77,* 4 - 20.

Smith, R., & Rennie, D. (2014). Evidence-based medicine – an oral history. *Journal of the American Medical Association, 311*, 365 – 367.

Sokal, M. M. (Ed.) (1987). *Psychological testing and American society: 1890 - 1930*. New Brunswick: Rutgers University Press.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology, 15,* 201 - 292.

Spearman, C. (1914). The heredity of abilities. *The Eugenics Review, 6,* 219 – 237.

Spearman, C. (1925). Some issues in the theory of "g" (including the law of diminishing returns). *Nature, 116,* 436 – 439.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* London: Macmillan.

Stein, Z. (2014). *Tipping the scales: Social justice and educational measurement*. Unpublished doctoral dissertation, Harvard Graduate School, Cambridge, MA.

Stephens, N., & Lewis, J. (2017). Doing laboratory ethnography: Reflections on method in scientific workplaces. *Qualitative Research, 17,* 202 – 2016.

Stern, A. M. (2005). *Eugenic Nation: Faults and Frontiers of Better Breeding in Modern America*. Berkeley: University of California Press.

Stichweh, R. (1992). The sociology of scientific disciplines: On the genesis and stability of the disciplinary structure of modern science. *Science in Context, 5,* 3 – 15.

Suppes, P. (1994). Ernest Nagel. November 16, 1901 – September 20, 1985. *Biographical Memoirs, 65,* 257 – 272.

Tenn, J. S. (2016). Introducing AstroGen: the Astronomy Genealogy Project. arXiv preprint:arXiv:1612.08908.

Terman, L. M. (1916). *The Measurement of Intelligence*. Boston, MA: Houghton Mifflin.

Terman, L. M., Kelley, T. L., & Ruch, G. M. (1922). *Stanford Achievement Test, Ed. 1922.* New York: World Book Company.

Thagard, P. (1999). *How scientists explain disease*. Princeton, NJ: Princeton University Press.

Thissen, D. (2001). Psychometric engineering as art. *Psychometrika, 66,* 473 – 485.

Thompson, G. (1947). Charles Spearman, 1863-1945. *Obituary Notices of Fellows of the Royal Society, 5*, 373 - 385.

Thompson, P. (2017). *The voice of the past: Oral history.* Oxford University Press.

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology, 8,* 271 – 281.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review, 34*, 273 – 286.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529 - 554.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38,* 406 - 427.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review, 41,* 1 – 32.

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: University of Chicago Press.

Thurstone, L. L. (1937). Psychology as a quantitative rational science. *Science, 85*, 227 - 232.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Thurstone, T. G. (1976). Dorothy C. Adkins. *Psychometrika, 41,* 434 – 437.

Irribarra, D. T. (2016). Ordering, measurement, and ordinal measurement: A pragmatic perspective. Unpublished doctoral dissertation, UC Berkeley.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16,* 8 – 14.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70,* 629 – 650.

Tversky, A. (1992). *Clyde Hamilton Coombs: July 22, 1912 – February 4, 1988. A biographical memoir*. Washington, DC: National Academy of Sciences.

Van Bork, R., Rhemtulla, M., & Borsboom, D. (2015). *Latent variable and network model implications for partial correlation structures.* Presentation for the 80th Annual Meeting of the Psychometric Society (IMPS), Beijing, July 2015.

Van Bork, R., Rhemtulla, M., Waldorp, L. J., & Borsboom, D. 2016. *Distinguishing latent variable models and network models*. Presentation for the 28th Annual Convention of the Association for Psychological Science (APS), Chicago, May 2016.

Van Bork, R., Wijsen, L. D., & Rhemtulla, M. (2017). Toward a causal interpretation of the Common Factor Model. *Disputatio, 9*, 581 – 601.

Van Borkulo, C. D., Borsboom, D., & Schoevers, R. A. (2016). Group-level symptom networks in depression—reply. *JAMA Psychiatry*, *73*, 411 – 412.

Van Borkulo, C. D., Boschloo, L., Borsboom, D., Penninx, B. W. J. H., Waldorp, L. J., & Schoevers, R. A. (2015). Association of Symptom Network Structure With the Course of Longitudinal Depression. *JAMA Psychiatry*, *72*, 1219 - 1226.

Van Fraassen, B. (2008). *Scientific representation: Paradoxes of perspective.* Oxford: University Press.

Van de Geer, J. P. (1971). *Introduction to multivariate analysis for the social sciences*. San Francisco, CA: Freeman.

Van der Heijden, P. G. M., & Sijtsma, K. (1996). Fifty years of measurement and scaling in the Dutch social sciences. *Statistica Neerlandica*, 50, 111 – 135.

Van der Linden, W. J., & Hambleton, R. K. (Eds.) (2013). *Handbook of modern item response theory.* New York: Springer.

Van der Maas, H. L. J., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*, 842 - 861.

Van Zandt, T., & Townsend, J. T. (2012). Mathematical psychology. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA Handbook of research methods in psychology, Vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological,* (pp. 369 - 386). Washington, DC: American Psychological Association.

Wainer, H. (2011). Profiles in research: An interview with Karl Gustav Jöreskog. *Journal of Educational and Behavioral Statistics, 36*, 403 – 412.

Weinberger, N. (2015). If intelligence is a cause, it is a within-subjects cause. *Theory & Psychology, 25*, 346 – 361.

Weiner, C. (1988). Oral History of Science: A Mushrooming Cloud? *The Journal of American History, 75,* 548 – 559.

What is psychometrics? (n.d.). Retrieved September 27, 2018, from http://psychometricsociety.org/

White, J. (2006). *Intelligence, destiny and education: The ideological roots of intelligence testing.* New York: Routledge.

Wicherts, J. M. (2007). Group differences in intelligence test performance. Unpublished doctoral dissertation, University of Amsterdam, Netherlands.

Wijsen, L. D., Borsboom, D., Cabaço, T., & Heiser, W. J. (2019). An academic genealogy of Psychometric Society presidents. *Psychometrika, 84,* 562 - 588.

Wijsen, L. D. (forthcoming). *Twenty interviews with Psychometric Society presidents: What drives the psychometrician?* Manuscript in preparation.

Williams, R. B. (1993). Contributions to the history of psychology: XCIII. Tracing academic psychology. *Psychological Reports, 72,* 85 – 86.

Wilson, M., & Gochyyev, P. (2013). Psychometrics. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 3 - 30). Rotterdam, The Netherlands: Sense Publishers.

Winston, A. S. (1998). Science in the service of the far right: Henry E. Garrett, the IAAEE, and the Liberty Lobby. *Journal of Social Issues, 54,* 179 – 210.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, UK: Oxford University Press.

Woodworth, R. S., & Thorndike, E. L. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8*, 247 – 261.

Wray, K. B. (2005). Rethinking scientific specialization. *Social Studies of Science, 35,* 151 – 164.

Wright, C., & Ville, S. (2017). The evolution of an intellectual community through the words of its founders: Recollections of Australia's economic history field. *Australian Economic History Review, 57,* 345 – 367.

Yanchar, S.C., & Hill, J. R. (2003). What is psychology about? Toward an explicit ontology. *Journal of Humanistic Psychology, 43,* 11 – 32.

Young, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing onexploratory analysis. *Tutorials in Quantitative Methods for Psychology, 9,* 79 - 94.

Young, F. W. (1996, June). New directions in psychology. Paper presented at the meeting of the Psychometric Society, Banff, Canada.

Appendix A

Table 2. Evidential Sources for each Advisor-Student Relationship in the Angell Genealogy.

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
Harvey A. Carr	University of Chicago	1905	James R. Angell	W. S. Hunter (1951). *James Rowland Angell (1869-1949), A biographical memoir.* Washington, DC: National Academy of Sciences.
Louis L. Thurstone	Chicago University	1917	James R. Angell	Dissertation
Carl J. Warden	University of Chicago	1922	Harvey A. Carr	Personal communication with Library University of Chicago.
Paul Horst	Chicago University	1931	Louis L. Thurstone	Heiser, W., & Hubert, L. (2016). A Creation Narrative for the Psychometric Society and Psychometrika: In the Beginning There Was Paul Horst. *Psychometrika*, *81*, 1172 - 1176.
Harold O. Gulliksen	University of Chicago	1931	Louis L. Thurstone	Personal communication with Robert Cudeck.
Robert L. Thorndike	Columbia University	1935	Carl J. Warden	Dissertation
Clyde H. Coombs	University of Chicago	1940	Louis L. Thurstone	Poole, K. T. (2008). The evolving influence of psychometrics in political science. In J.M. Box-Sfeffensmeier, H. E. Brady & D. Collier (Eds.) *The Oxford handbook of political methodology* (pp. 199-216). Oxford: University Press.
Ledyard R. Tucker	University of Chicago	1946	Louis L. Thurstone	Dorans, N. J. (2004). *A conversation with Ledyard R. Tucker.* ETS Publication.
Warren S. Torgerson	Princeton	1951	Harold O. Gulliksen	Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 17,* 401 – 419.
Bert F. Green, Jr.	Princeton University	1951	Harold O. Gulliksen	Dissertation
Frederic M. Lord	Princeton University	1951	Harold O. Gulliksen	Dissertation
Samuel J. Messick	Princeton University	1954	Harold O. Gulliksen	Dissertation
Norman Cliff	Princeton	1956	Harold O. Gulliksen	Personal communication with Robert Cudeck.
William M. Meredith	University of Washington	1958	Paul Horst	Dissertation

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
J. Douglas Carroll	Princeton University	1963	Harold O. Gulliksen	Dissertation
Robert Linn	University of Illinois	1965	Ledyard R. Tucker	Personal communication with Bill Stout.
Bruce Bloxom	University of Washington	1966	Paul Horst	Bloxom, B. (1967). Effects of anger-arousing instructions on personality questionnaire performance. *ETS Research Report Series, 1,* 1-14.
James O. Ramsay	Princeton University	1966	Harold O. Gulliksen	Personal communication with James Ramsay.
Forrest W. Young	University of Southern California	1967	Norman Cliff	Personal communication with Yoshio Takane.
Philip L. Smith	University of Illinois	1976	Robert Linn	Personal communication with Terry Ackerman.
Yoshio Takane	University of North Carolina at Chapel Hill and University of Tokyo	1977	Forrest W. Young	Personal communication with Yoshio Takane.
Robert Cudeck	University of Southern California	1980	Norman Cliff	Personal communication with Robert Cudeck.
Roger E. Millsap	University of California	1983	William M. Meredith	Maydeu-Olivares, A. (2014). In Memoriam, Roger E. Millsap 1954-2014. *Psychometrika 79,* 355 - 356.
Terry A. Ackerman	University of Wisconsin-Milwaukee	1984	Philip L. Smith	Personal communication with Terry Ackerman.

Table 3. Evidential Sources for each Advisor-Student Relationship in the Wundt Genealogy.

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
Hugo Münsterberg	University of Leipzig	1885	Wilhelm Wundt	Hilgard, E. R. (1987). *Psychology in America: A historical survey.* Orlando, FL: Harcourt, Brace, Jovanovich.
James McKeen Cattell	Leipzig Univeristy	1886	Wilhelm Wundt	Hilgard, E. R. (1987). *Psychology in America: A historical survey.* Orlando, FL: Harcourt, Brace, Jovanovich.
Robert S. Woodworth	Columbia University	1889	James McKeen Cattell	Boring, M. D., & Boring E. G. (1948). Masters and pupils among the American psychologists. *The American Journal of Psychology, 61,* 527 – 534.
Edward B. Titchener	University of Leipzig	1892	Wilhelm Wundt	Hilgard, E. R. (1987). *Psychology in America: A historical survey.* Orlando, FL: Harcourt, Brace, Jovanovich.
Charles H. Judd	University of Leipzig	1896	Wilhelm Wundt	Buswell, G. T. (1947). Charles Hubbard Judd: 1873-1946. *The American Journal of Psychology, 60,* 135 – 137.
Edward L. Thorndike	Columbia University	1898	James McKeen Cattell	Hilgard, E. R. (1987). *Psychology in America: A historical survey.* Orlando, FL: Harcourt, Brace, Jovanovich.
Robert M. Yerkes	Harvard University	1902	Hugo Münsterberg	Hilgard, E. R. (1987). *Psychology in America: A historical survey.* Orlando, FL: Harcourt, Brace, Jovanovich.
Charles E. Spearman	Leipzig University	1906	Wilhelm Wundt	Williams, R. H., Zimmerman, D. W., Zumbo, B. D., & Ross, D. (2003). Charles Spearman: British Behavioral Scientist. *Human Nature Review, 3,* 114 – 118.
Herbert H. Woodrow	Columbia University	1909	Robert S. Woodworth	Dissertation
Melvin E. Haggerty	Harvard University	1910	Robert M. Yerkes	Haggerty, Melvin, E. (Melvin Everett), 1875-1937. (n.d.). Retrieved from http://snaccooperative.org/ark:/99166/w6f197qx
Albert T. Poffenberger	Columbia University	1912	Robert S. Woodworth	Schoenfeld, W. N. (1979). Albert Theodore Poffenberger: 1885 – 1977. *The American Journal of Psychology, 92,* 143 – 149.
Truman L. Kelley	Columbia University	1914	Edward L. Thorndike	Hubert, L. (2013). Truman Lee Kelley (1884-1961) [pdf]. Retrieved from http://cda.psych.uiuc.edu/kelley_beamer_talk.pdf
Guy T. Buswell	University of Chicago	1920	Charles H. Judd	California Digital Library: http://texts.cdlib.org/view?docId=hb5g50061q&doc.view=frames&chunk.id=div00019 ;

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
Herbert A. Toops	Columbia University	1921	Edward L. Thorndike	Koppes, L. L. (2014). *Historical perspectives in industrial and organizational psychology.* Mahwah, NJ: Lawrence Erlbaum.
Henry E. Garrett	Columbia University	1922	Albert T. Poffenberger	Dissertation
Karl John Holzinger	University of Chicago	1922	Charles E. Spearman	Sternberg, R. J., & Grigorenko, E. L. (Eds). (2002). *The general factor of intelligence: How general is it?* Mahwah, NJ: Lawrence Erlbaum.
Adam R. Gilliland	University of Chicago	1922	Charles H. Judd	Annotated List of Ph.D. Dissertations in Reading, 1916 – 1969. Chicago University.
Joy P. Guilford	University of Illinois	1927	Edward Titchener	Brett, J. M., & Drasgow, F. (Eds.). (2002). *The psychology of work: Theoretically based empirical research.* London: Lawrence Erlbaum.
Harold A. Edgerton	Ohio State University	1928	Herbert A. Toops	Edgerton, H. A. (n.d.). A career in industrial and measurement psychology. Retrieved from http://www.siop.org/presidents/Edgerton.aspx
Robert. J. Wherry	Ohio State University	1929	Herbert A. Toops	Bartlett, C. J. (1982). The legacy of Robert J. Wherry, Sr. (1904-1981). *The Industrial-Organizational Psychologist, 5,* 7.
Irving Lorge	Teachers College, Columbia	1930	Edward L Thorndike	Thorndike, R. L. (1961). Irving Lorge. *Psychometrika, 26,* 1-2.
Jack W. Dunlap	Columbia University	1931	Edward L. Thorndike	Benjamin, L. T. (n.d.). The Early Presidents of Division 14: 1945-1954. Retrieved from http://www.siop.org/tip/backissues/tipoct97/BENJAM~1.aspx
Edward E. Cureton	Teachers College, Columbia	1931	Edward L. Thorndike	Shrader, R. R. (1994). Edward E. (Ted) Cureton (1902-1992). *American Psychologist, 49,* 350.
Philip J. Rulon	University of Minnesota	1931	Melvin E. Haggerty	Dissertation
Philip H. Dubois	Columbia University	1932	Henry E. Garrett	Thumin, F. J., & Barclay, A. G. (2002). Philip Hunter Dubois (1903 – 1998). *American Psychologist, 57,* 368.
John C. Flanagan	Harvard University	1935	Truman L. Kelley	Hubert, L. (2013). Truman Lee Kelley (1884-1961) [pdf]. Retrieved from http://cda.psych.uiuc.edu/kelley_beamer_talk.pdf
Dorothy C. Adkins	Ohio State University	1937	Herbert A. Toops	Adkins, Dorothy Christina, 1912-1975. Retrieved from http://snaccooperative.org/ark:/99166/w6hx2jvc

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
Hubert E. Brogden	University of Illinois	1939	Herbert H. Woodrow	Dissertation
Lee J. Cronbach	University of Chicago	1940	Guy T. Buswell	Shavelson, L. J. (2009) *Lee J. Cronbach (1916-2001), A biographical memoir.* Washington, DC: National Academy of Sciences.
Allen E. Edwards	Northwestern University	1940	Adam R. Gilliland	Summaries of Doctoral Dissertations, Northwestern University.
Ernest A. Haggard	Harvard University	1946	Truman L. Kelley	Personal communication with Harvard University Archives
Chester W. Harris	University of Chicago	1946	Karl J. Holzinger	Dissertation
Ben J. Winer	Ohio State University	1951	Robert J. Wherry	*Distinguished Teaching of Quantitative Methods in Psychology Award* [PDF file]. *American Psychologist, 39,* 313 – 314. Retrieved from: https://www.psychometricsociety.org/sites/default/files/Ben_J_Winer_from_Purdue.pdf
R. Darrell Bock	University of Chicago	1952	Ernest A. Haggard	Personal communication with R. Darrell Bock
Shizuhiko Nishisato	University of Chicago	1965	R. Darrell Bock	CV R. Darrell Bock
David Thissen	University of Chicago	1976	R. Darrell Bock	CV R. Darrell Bock
Robert Mislevy	University of Chicago	1981	R. Darrell Bock	Personal communication with Robert Mislevy
Ulf Böckenholt	University of Chicago	1985	R. Darrell Bock	CV R. Darrell Bock
Albert Maydeu-Olivares	University of Illinois	1997	Ulf Böckenholt	Personal Communication with Albert Maydeu-Olivares.

Table 4. Evidential Sources for each Advisor-Student Relationship in the James Genealogy.

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
G. Stanley Hall	Harvard University	1878	William James	Hilgard, E. R. (1987). *Psychology in America: A historical survey.* Orlando, FL: Harcourt, Brace, Jovanovich.
Joseph Jastrow	John Hopkins University	1886	G. Stanley Hall	History of psychology. Retrieved from: http://www.learner.org/series/discoveringpsychology/history/history_nonflash.html
Lewis M. Terman	Clark University	1905	G. Stanley Hall	Boring. E. G. (1959). *Lewis Madison Terman (1877 – 1856), A biographical memoir.* Washington, DC: National Academy of Sciences.
Morris R. Cohen	Harvard University	1906	William James	Personal communication with Harvard University Library
Clark L. Hull	University of Wisconsin	1918	Joseph Jastrow	Rieber, R. (2012). *Encyclopedia of the history of psychological theories*. New York: Springer.
Ernest Nagel	Columbia University	1931	Morris R. Cohen	Mathematics Genealogy Project
Quinn McNemar	Stanford University	1932	Lewis M. Terman	Hastorf, A. H. (2004). *History: Illustrious from past to present.* Retrieved from: https://psychology.stanford.edu/about/history
Neal E. Miller	Yale University	1935	Clark L. Hull	Coons, E. E. (2014). *Neal E. Miller (1909-2002), A biographical memoir.* Washington, DC: National Academy of Sciences.
Carl I. Hovland	Yale University	1936	Clark L. Hull	Hurley, K. P., & Hogan, J. D. (2017, June). *Carl Iver Hovland: A model general psychologist. A spotlight on Past-Presidents of APA Div. 1.* Retrieved from: http://www.apadivisions.org/division-1/publications/newsletters/general/2017/06/hovland-profile.aspx
Patrick C. Suppes	Columbia University	1950	Ernest Nagel	Mathematics Genealogy Project
Roger N. Shepard	Yale University	1955	Carl I. Hovland	Shepard, R. N. (1998). Carl Iver Hovland June 12, 1912 – April 16, 1961. *Biographical memoirs. National Academy of Sciences, 73,* 231 – 261.
Gordon H. Bower	Yale University	1959	Neal E. Miller	Chamberlin J. (2007). Psychologist wins National Medal of Science. *Monitor on Psychology, 38,* 10.
Paul W. Holland	Stanford University	1966	Patrick C. Suppes	Personal Communication with Paul W. Holland
Lawrence J. Hubert	Stanford University	1971	Patrick C. Suppes	Personal Communication with Larry Hubert
Phipps Arabie	Stanford University	1974	Gordon H. Bower	Personal communication Willem J. Heiser

Table 5. Evidential Sources for each Advisor-Student Relationship in the Michotte Genealogy.

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
Franciscus J. M.A. Roels	Université Catholique de Louvain	1914	Albert E. Michotte	Busato, V., Essen, M. van., & Koops, W. (Eds.) (2013). *Vier grondleggers van de psychologie*. Amsterdam: Bert Bakker.
F.J. Theo Rutten	Utrecht University	1929	Franciscus J. M. A. Roels	Busato, V., Essen, M. van., & Koops, W. (Eds.) (2016). *Zeven grondleggers van de psychologie*. Amsterdam: Bert Bakker.
Alfons M. J. Chorus	Katholieke Universiteit Nijmegen	1940	F.J. Theo Rutten	Busato, V., Essen, M. van., & Koops, W. (Eds.)(2016). *Zeven grondleggers van de psychologie*. Amsterdam: Bert Bakker.
Jozef R. Nuttin	Catholic University Leuven	1941	Albert E. Michotte	Avermaet, E. van. (n.d.). JR Nuttin: Biography. Retrieved from https://ppw.kuleuven.be/home/english/faculty/jrnuttin/biography
John P. van de Geer	Leiden University	1957	Alphons M.J. Chorus	Busato, V., Essen, M. van., & Koops, W. (Eds.) (2016). *Zeven grondleggers van de psychologie*. Amsterdam: Bert Bakker.
Willem Claeys	Catholic University Leuven	1963	Jozef R. Nuttin	Personal communication with Paul De Boeck
Jan de Leeuw	Leiden University	1973	John P. van de Geer	Busato, V., Essen, M. van., & Koops, W. (Eds.) (2016). *Zeven grondleggers van de psychologie*. Amsterdam: Bert Bakker.
Paul De Boeck	Catholic University Leuven	1977	Willem Claeys	Personal communication with Paul De Boeck
Jos M. F. ten Berge	University of Groningen	1977	John P. van de Geer	Busato, V., Essen, M. van., & Koops, W. (Eds.) (2016). *Zeven grondleggers van de psychologie*. Amsterdam: Bert Bakker.
Willem J. Heiser	Leiden University	1981	John P. van de Geer	Busato, V., Essen, M. van., & Koops, W. (Eds.) (2016). *Zeven grondleggers van de psychologie*. Amsterdam: Bert Bakker.
Jacqueline J. Meulman	Leiden University	1986	John P. van de Geer	Busato, V., Essen, M. van., & Koops, W. (Eds.) (2016). *Zeven grondleggers van de psychologie*. Amsterdam: Bert Bakker.
Francis Tuerlinx	Catholic University Leuven	2000	Paul De Boeck	Personal communication with Francis Tuerlinckx

Table 6. Evidential Sources for each Advisor-Student Relationship in Table 1.

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
John R. Knott	University of Iowa		Lee Edward Travis	Magoun H. W. (2003). American neuroscience in the twentieth century. Lisse, The Netherlands: A. A. Balkema Publishers.
Georg Elias Müller	Gottingen University	1873	Hermann Lotze	Boring, E. G. (1935). Georg Elias Müller: 1850-1934. *The American Journal of Psychology, 47,* 344 – 348.
Carl Emil Seashore	Yale University	1895	George Trumbull Ladd	Miles, W. R. (1956). *Carl Emil Seashore (1866-1949), A biographical memoir.* Washington, DC: National Academy of Sciences.
Raymond Dodge	Halle University	1896	Benno Erdmann	Miles, W. R. (1956). *Raymond Dodge (1871 – 1942), A biographical memoir.* Washington DC: National Academy of Sciences.
Erich Becher	University of Bonn	1904	Benno Erdman	Personal communication with Gerhard Fischer
Géza Révész	Gottingen University	1905	Georg Elias Müller	Levelt, W. J. M. (2013). *A history of psycholinguistics: The pre-Chomskyan era.* Oxford: University Press.
Stuart Chapin	Columbia University	1911	Franklin H. Giddings	Platt, J. (1998). *A history of sociological research methods in America, 1920-1960.* Cambridge University Press.
Lee Edward Travis	University of Iowa	1924	Carl Emil Seashore	Miles, W. R. (1956). *Carl Emil Seashore (1866-1949), A biographical memoir.* Washington, DC: National Academy of Sciences.
Hubert Rohracher	University of Munich	1926	Erich Becher	Personal communication with Gerhard Fischer
Harold Dean Carter	University of Minnesota	1930	Donald G. Paterson	Personal communication with University of Minnesota Libraries
B. F. Skinner*	Harvard University	1931	William John Crozier	Mathematics Genealogy Project
John M. Hadley	University of Iowa	1935	John R. Knott	Personal communication Special Collections University of Iowa.
Bruno Bettelheim	University of Chicago	1937	Robert Reininger	Pollak. R. (1997). *The creation of Doctor B: A biography of Bruno Bettelheim.* New York: Touchstone Books.
Lloyd Humphreys	Stanford University	1938	Ernest Hilgard	Lubinski, D. (2004). Lloyd G. Humphryes: Quintessential Scientist (1913-2003). *Intelligence, 32,* 221 – 226.
Ernest Hilgard	Yale University	1940	Raymond Dodge	Bower, G. H. (2010). *Ernest Ropiequet Hilgard (1904 – 2001), A biographical memoir.* Washington, DC: National Academy of Sciences.

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
John B .Carroll	University of Minnesota	1941	Burrhus Frederic Skinner	Lubinski, D. (2004). John Bissell Carroll (1916-2003). *American Psychologist, 59,* 43-33.
Louis Guttman	University of Minnesota	1942	Stuart Chapin	Katz, E. Louis Guttman, 1916-1987. *Public Opinion Quarterly, 52,* 240 – 242.
Adriaan D. Groot	Universiteit van Amsterdam	1946	Géza Révész	Dissertation
Erich Mittenecker	Vienna University	1947	Hubert Rohracher	Personal communication with Gerhard Fischer
Hendrik S. Steyn*	University of Edinburgh	1949	Alexander Craig Aitken	Mathematics Genealogy Project
R. Duncan Luce*	Massacusetts Institute of Technology	1950	Irvn S. Cohen	Mathematics Genealogy Project
Lyle V. Jones	Stanford University	1950	Lloyd G. Humphreys	Thissen, D. (2016). Remembering Lyle V. Jones (1924 – 2016). *Psychometrika, 81,* 904 – 905.
Douglas N. Jackson	Purdue University	1955	John M. Hadley	Cautin, R. L., & Lilienfeld, S. O. (Eds). (2015). *The Encyclopedia of Clinical* Psychology, Volumr 5. Maldon, MA: John Wiley & Sons.
Rene Dawis	University of Minnesota	1956	Donald G. Paterson	Personal communication
Henry F. Kaiser	University of California, Berkeley	1956	Harold Dean Carter	Kaiser, H. R. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187 – 200.
Benjamin D. Wright	University of Chicago	1957	Bruno Bettelheim	Rasch Measurement Transactions, Volume 29:3, Winter 2015.
Gerhard Fischer	Vienna University	1963	Erich Mittenecker	Personal communication with Gerhard Fischer
Peter Bentler	Stanford University	1964	Douglas N. Jackson	Personal communication
Fumiko Samejima	Keio University	1965	Taro Indow	Wainer, H., & Robinson, D. H. (2007). Profiles in Research: Fumiko Samejima. *Journal of Educational and Behavioral Statistics, 32,* 206 – 222.
Roderick Mcdonald	University of Queensland	1967	John Keats	McDonald, R. P., Maydeu-Olivares, A., & McArdle, J. J. (2005). Contemporary psychometrics: A festschrift for Roderick P. McDonald. Mahwah, NJ: Lawrence Erlbaum.

Name of Scholar	University of Graduation	Year of Graduation	Doctoral Advisor	Source
Michael W. Browne	University of South Africa	1968	Hendrik S. Steyn	Browne, M. W. (1969). *Precision of prediction* (Research Bulletin No. 69-69). Princeton, NJ: Educational Testing Service.
Ivo W. Molenaar*	University of Amsterdam	1970	Jan Hemelrijk	Mathematics Genealogy Project
Gideon Mellenbergh	University of Amsterdam	1971	Adriaan D. de Groot	Personal communication with Paul De Boeck
Susan Embretson	University of Minnesota	1973	Rene Dawis	Personal communication
Wim van der Linden	University of Amsterdam	1980	Gideon Mellenbergh	Personal communication with Willem van der Linden
Mark R. Wilson	University of Chicago	1984	Benjamin D. Wright	Avermaet, E. van. (n.d.). JR Nuttin: Biography. Retrieved from https://ppw.kuleuven.be/home/english/faculty/jrnuttin/biography
Klaas Sijtsma	University of Groningen	1988	Ivo W. Molenaar	Personal communication with Klaas Sijtsma
Cees Glas	University of Twente	1989	Willem J. van der Linden	University website
Sophia Rabe-Hesketh	King's College, University of London	1992	James F. Boyce	Personal communication with Sophia Rabe-Hesketh
Anders Skrondal	University of Oslo	1996	Petter Laake	Personal communication with Anders Skrondal

* The individual lineages of these scholars can be found in its entirety on the website of the Mathematics Genealogy Project.

Appendix B

Questions for interviews (Chapter 3)

Introduction text: Thank you for your participation in this oral history project on the history of psychometrics. In this interview, I will be asking questions on your career as a psychometrician, the relation between psychology and psychometrics, and of course, your view on the history and future of psychometrics. Well, let's start.

1. The career of the interviewee

1. How did you end up in psychometrics? What did you originally study?
2. Did you always want to become a researcher? Have you ever considered switching to another career entirely?
3. Is there a specific person who sparked your passion for psychometrics? Who was this?
4. Who were your supervisors?
5. Can you give me a short overview of your career so far? Where did you start your Ph.D.?
6. Looking back, can you identify the three most important research topics in your career so far?
7. When you consider your strongest critics, what is their strongest point of criticism? And their second strongest point?
8. Have you ever had considerable doubt about the direction you took in your research? Did you ever think that your critics were possibly right?
9. Have you ever questioned your own research?

2. The relation between psychology and psychometrics.

Psychology and psychometrics are of course two related fields, but have in some ways become quite distinct. For example, the psychometric society is solely aimed at psychometrics, not at psychology, and psychology and psychometrics have different institutions.

1. What do you think is the current relation of psychometrics and psychology?
   What do you think should be the relation between psychometrics and psychology?
2. Should psychology and psychometrics even be two separated disciplines?
3. What can psychometrics contribute to psychology, and the other way around?
4. Do you consider yourself to be a psychologist, or do you perhaps sympathize more with other fields (e.g. statistics, computational science, mathematics)?
5. What part of your research is still strongly rooted in psychology?

6. Do you consider your research to be interdisciplinary? What other scientific disciplines can we trace in your research?
7. What is the role of psychometrics in society? What is psychometrics' greatest contribution to society?

3. The history and future of psychometrics

8. What is your most cited paper?
9. When you look back at your career so far, what do personally think is your most influential or best article you've written?
10. What do you think is your legacy? What will you be remembered for?
11. What do you believe is the most important article or book ever written in psychometrics? Why?
12. Who is according to you the most important psychometrician that ever lived? Why?
13. How have the above influenced your own work?
14. What is psychometrics' biggest achievement?
15. What do you think is psychometrics' biggest challenge?
16. What do you think is the field that psychometrics can learn from the most? Why?/ Do you envy other fields?

In case the person is still working:

17. What is your personal challenge/goal for the coming years?
18. Is there something or someone that you can still learn from?

Is there anything you want to add to the interview? Something you want to respond to?
What did you think of this interview?

Nederlandse Samenvatting

De psychometrie is een wetenschappelijk vakgebied dat zich bezighoudt met het meten van psychologische attributen, maar echter wel op een zeer abstracte wijze. Psychometrici (met name zij die centraal staan in dit proefschrift) schrijven geen toets items, noch ontwikkelen zij theorieën over psychologische attributen. Wat zij wel doen is het ontwikkelen van psychometrische of statistische modellen. Deze modellen specificeren vaak de relatie tussen latente variabelen, zoals cognitieve vaardigheden, en observeerbare variabelen, zoals de antwoorden op toets items. Velen van ons zijn bekend met de toepassingen van de psychometrie, zoals *job assessments*, het leerlingvolgsysteem van CITO voor basisschoolleerlingen, de SAT en ACT toetsen voor toelating op Amerikaanse universiteiten, en diagnostische meetinstrumenten voor psychologische stoornissen. De psychometrie als een wetenschappelijk vakgebied is echter moeilijk te begrijpen voor mensen die zelf geen psychometricus zijn. De complexiteit van het vakgebied roept veel vragen op over waar de psychometrie nou werkelijk over gaat, zijn historische ontwikkeling en zijn wetenschapsfilosofie. Wat doen de psychometrici nou eigenlijk, en wat vinden ze belangrijk? Wat voor modellen gebruiken ze en wat impliceren deze modellen over psychometrisch onderzoek? Hoe heeft het vakgebied zich historisch ontwikkeld? Dit proefschrift probeert door de toepassing van verschillende methoden (de etnografische methode, de historische methode en de wetenschapsfilosofische methode) verschillende karakterschetsen van de psychometrie op te tekenen. Elk van deze karakterschetsen richt zich op een specifiek aspect van de mysterieuze psychometrie.

Een historische karakterschets

Hoofdstuk 2 visualiseert een deel van de geschiedenis van de psychometrie aan de hand van een aantal academische genealogieën of stambomen van presidenten van de *Psychometric Society*. De meerderheid van de presidenten blijkt af te stammen van invloedrijke psychologen als Wilhelm Wundt, William James, James Angell, of Albert Michotte. Een kleinere doch aanzienlijke groep stamt af van wiskundige Carl F. Gauss. De stambomen hebben naast een beschrijvende functie, ook het doel om belangrijke historische kennis, zoals de connecties tussen wetenschappers en hun begeleiders, te behouden. Op basis van de stambomen hebben we een aantal interessante aspecten van de geschiedenis van de psychometrie geformuleerd. Als eerste is het frappant dat een aantal invloedrijke wetenschappers, zoals Francis Galton en Gustav Fechner, niet in de stambomen zijn opgenomen, en andere invloedrijke wetenschappers (zoals Charles Spearman) slechts één of twee navolgers hebben. Er zijn een aantal redenen waarom dit het geval kan zijn, maar wat het in ieder geval aantoont is dat het succes van een psychometricus niet hand in hand gaat met het aantal academische afstammelingen. Van

geschiedkundig belang zijn aan de ene kant, en een rol spelen in de institutionalisering van de psychometrie als wetenschappelijk vakgebied of de *Psychometric Society* aan de andere kant, zijn dus relatief onafhankelijke factoren. Ten tweede tonen de stambomen de fragmentatie van de psychometrie; het is duidelijk zichtbaar dat er de psychometrie is geworteld in zowel de psychologie als in de wiskunde. Een suggestie die we hier maken is dat het multidisciplinaire karakter van de psychometrie verantwoordelijk is voor de spanning in de psychometrie tussen onderzoekers die psychologie-georiënteerd zijn, en onderzoekers die meer statistiek-georiënteerd zijn. Deze spanning heeft zich in de loop der tijd geuit door verschillende schisma's tussen de *Psychometric Society* en andere takken in de kwantitatieve psychologie. Ten derde tonen de stambomen aan dat sinds het begin van de Psychometric Society in 1937, de psychometrie diverser is geworden zowel in termen van gender, als in termen van wetenschappelijke en socio-culturele achtergrond van de psychometrici. Hier doen we de suggestie dat de *Psychometric Society* nu minder de psychometrische gemeenschap vertegenwoordigt dan in zijn vroege jaren, toen de gemeenschap beperkt was tot slechts een aantal universiteiten.

De historische karakterschets heeft een vervolg in hoofdstuk 6, waarin we een vergelijking maken tussen het gebruik van waarden in de vroege psychometrie en het gebruik van waarden in de moderne psychometrie. Verschillende psychometrie uit de vroege jaren waren expliciet toegewijd aan een eugenetische ideologie en een nieuwe sociale orde, waarin de maatschappij was geordend op basis van cognitieve vaardigheid (intelligentie). Mensen met een hoge intelligentie moesten terecht komen in zeer veeleisende posities, zoals die van politiek leider, en mensen met een lage intelligentie waren voorbestemd om minder gewaardeerd manueel werk te doen. En aangezien werd aangenomen dat intelligentie erfelijk was, zou de ideale samenleving mensen met een hoge intelligentie aanmoedigen zich voort te planten, en mensen met een lage intelligentie dit ontmoedigen. Het objectief meten van psychologische attributen moest een belangrijke methode worden een rol zou spelen in het ordenen van mensen op basis van hun intelligentie en het toeschrijven van geschikte taken. Een beroemd voorbeeld hiervan is het testen van militairen tijdens de Eerste Wereldoorlog, toen psychologische tests gebruik werden om onderscheid te maken tussen geschikte en ongeschikte soldaten. De psychometrici waren dus expliciet over welke waarden hun onderzoek beïnvloedden. Zoals Hoofdstuk 6 aantoont, is de moderne psychometrie minder expliciet over de waarden die zij hanteert, al zijn deze echter wel aanwezig. Aangezien de analyse over waarden in de moderne psychologie meer getuigt van een wetenschapsfilosofische aanpak, zal ik hier over uitweiden bij de filosofische karakterschets (zie hieronder).

Een etnografische karakterschets

De psychometrici vormen een groep met een specifieke cultuur van onderzoeksinteressen, drijfveren en praktijken. Door middel van een *oral history* studie, hebben we onderzocht hoe deze psychometrici reflecteren op on hun eigen vakgebied. Het doel van deze studie was het verkrijgen van een beter begrip van de belangrijkste historische, moderne en mogelijk toekomstige aspecten die een rol spelen in de psychometrie, en met name ook begrip over hoe de psychometrici op deze aspecten reflecteren. Gebaseerd op de interviews, kunnen we concluderen dat de psychometrie een veelzijdig vakgebied is. Sommige psychometrici zijn meer psychologie-georiënteerd, en vinden dat de psychometrie zou moeten dienen voor psychologische theorievorming. Anderen zijn meer statistiek- of data-analyse-georiënteerd. Deze groep beschouwt de psychometrie als een specifieke groep statistische modellen die in principe kunnen worden getransporteerd naar andere onderzoeksgebieden. Naast de psychologie-oriëntatie en de statistiek-oriëntatie, vonden we ook de consultatie-, de ingenieurs-, en de wiskundige oriëntatie. Psychometrici hebben dus verschillende perspectieven over wat voor aanpak leidend moet zijn in psychometrisch onderzoek.

Behalve de diversiteit in psychometrisch onderzoek, blijkt ook dat de psychometrici divers reflecteren op de toekomst van hun vakgebied. Sommigen zijn optimistisch over een toekomst voor de psychometrie, ofwel omdat toetsing in onderwijs nog steeds een belangrijke rol speelt in de maatschappij, ofwel omdat de psychometrie veel te bieden heeft aan en te leren heeft van nieuwe ontwikkelingen als *data mining*, *big data* en *machine learning*. Anderen zijn meer op hun hoede wat betreft een vruchtbare toekomst voor de psychometrie. Als een vakgebied heeft de psychometrie last gehad van communicatieproblemen. Een zorg voor deze psychometrici is namelijk dat de psychometrie zichzelf slecht kan verkopen, zowel aan de psychologie, als aan de statistiek. Psychologen zijn minder geneigd om psychometrische literatuur te lezen, en de statistici voelen niet de noodzaak om bekend te zijn met een klein vakgebied als de psychometrie. De moderne psychometrie is dus een veelzijdig vakgebied dat zich meer dient te verdiepen in de verbinding met andere, gerelateerde vakgebieden. Gebaseerd op de interviews, raden we aan dat de psychometrie zijn veelzijdigheid dient te erkennen en moet specificeren wat het wil bereiken in de toekomst.

Een filosofische karakterschets

De derde en laatste karakterschets is de filosofische karakterschets, die in twee delen kan worden opgesplitst. Het eerste deel bestaat uit een causale lezing van psychometrische modellen (zie Hoofdstuk 4 en 5). Veel psychometrische modellen (o.a. reflectieve, formatieve en netwerkmodellen) impliceren een specifieke causale structuur, wat deze modellen geschikte instrumenten maakt voor psychologische theorievorming. Deze

hoofdstukken illustreren wat het betekent om een causale lezing te geven van deze modellen en waarom een causale lezing bruikbaar en geschikt is in bepaalde contexten. Bijvoorbeeld, het *common factor* model kan gebruikt worden als een causaal model als men kan aannemen dat de data een gelijkvormige structuur impliceert. Bovendien, een causale interpretatie van dit model legitimeert de focus op gedeelde in plaats van de unieke variantie. Daarnaast legitimeert een causale interpretatie van het *common factor* model de aanname van lokale onafhankelijkheid. De causale lezing van psychometrische modellen benadrukt dat veel van deze modellen causale relaties tussen variabelen impliceren, en dat ze gebruikt kunnen worden voor verklaring. Het is echter wel belangrijk om te realiseren dat causaal gebruik van psychometrische modellen gepaard gaat met een aantal controverses, zoals de causale status van individuele verschillen, *interpretational confounding* en het probleem van generalisatie (zie Hoofdstuk 4).

De causale lezing van psychometrische modellen geeft aan dat deze modellen de potentie hebben voor verklaring: de modellen kunnen worden gebruikt om verklaringen te formuleren voor menselijk gedrag. Echter, deze lezing is niet noodzakelijkerwijs een goede beschrijving van hoe psychometrische modellen ook werkelijk worden gebruikt. Het tegenovergestelde is wellicht waarschijnlijker: de causale of realistische interpretatie van psychometrische modellen staat in contrast met de meer data-analytische aanpak die populair is onder moderne psychometrici.

Het tweede deel van de filosofische karakterschets betreft een analyse over het gebruik van waarden in de moderne psychometrie (zie Hoofdstuk 6). Zoals eerder al genoemd, was de vroege psychometrie (late 19^{e} eeuw/vroege 20^{e} eeuw) expliciet toegewijd aan sociale en politieke waarden uit die tijd. De moderne psychometrie lijkt echter een discipline die relatief vrij van waarden is, aangezien zij zich vooral bezighoudt met zeer technische problemen. We laten zien dat ook de zeer technische moderne psychometrie een aantal specifieke waarden onderschrijft. Een voorbeeld is dat in de psychometrie een zeer specifiek idee over eerlijkheid wordt aangehangen dat als volgt kan worden omschreven: toetsen en items zijn pas eerlijk wanneer respondenten met dezelfde vaardigheid uit verschillende groepen dezelfde kansen hebben voor het correct beantwoorden van elk item. Eerlijkheid zoals geformuleerd door de psychometrici wijkt sterk af van andere noties van eerlijkheid, zoals economische of politieke eerlijkheid. Daarmee is het een waarde die specifiek is voor de psychometrie. Andere waarden die door de moderne psychometrie worden aangehangen zijn de conceptualisatie van individuele verschillen als kwantitatieve (niet kwalitatieve) verschillen, het nastreven van objectief meten (persoonlijke voorkeuren mogen het meetproces niet in de weg zitten) en de voorkeur voor nut in plaats van waarheid (in plaats van een interesse hebben in het waarheidsgehalte van psychologische theorie, zijn psychometrici vaak vooral gedreven door het praktisch nut van hun onderzoek).

Een portret van de psychometrie

Vanuit het perspectief van een buitenstaander, lijkt de psychometrie een mysterieuze, ingewikkelde en vrij uniforme wetenschap. Dit proefschrift had als doel om een aantal complexe aspecten van de psychometrie te belichten en toegankelijk te maken. De drie karakterschetsen hierboven vormen samen een multidimensionaal portret van de psychometrie. Psychometrie blijkt een multidisciplinair en divers vakgebied te zijn, dat wordt gekarakteriseerd door een aantal methodologische aanpakken en met een oorsprong in zowel de psychologie als de wiskunde. Dit proefschrift toont ook aan dat de psychometrie wordt gekenmerkt door een aantal interne spanningen. De psychometrie bevindt zich op het grensvlak tussen zijn toewijding aan psychologie of aan de statistiek, tussen het uitdragen van een sociale missie of zich te identificeren als technische discipline en tussen het gebruiken van modellen die geschikt zijn voor psychologische theorievorming of zich te richten op het bouwen van bruikbare toepassingen. In conclusie hoop ik dat dit proefschrift een aantal complexe eigenschappen van de psychometrie heeft belicht en de psychometrie toegankelijker heeft weten te maken voor mensen buiten de psychometrische gemeenschap.

English Summary

Psychometrics is a scientific discipline which concerns itself with psychological measurement, but in a mostly abstract fashion. Rather than writing test items and theorizing about measurable constructs, psychometricians (at least those who occupy the center stage in this dissertation) often concern themselves with the development of psychometric or statistical models. These models often specify a measurement relationship between unobservable variables, like cognitive abilities, and observable variables, like item responses. Though the applications of psychometrics are often quite familiar to many people (examples are job assessments, the CITO student tracking system test for Dutch school children, the SAT and ACT tests for college admissions in the United States, and diagnostic assessments for psychological disorders), psychometrics as a scientific discipline is relatively unintelligible for those who are not part of the inner circle. The impenetrable character of psychometrics raises many questions about the practice of psychometric research, its historical development, and its philosophy of science. What is it that psychometricians do and what do they actually care about? How has psychometrics developed historically? What kind of models do psychometricians work with and what do these models imply about psychometric research? By using a number of approaches – namely that of the historian, that of the ethnographer, and that of the philosopher of science – this dissertation aims to draw different characterizations of psychometrics, each highlighting different aspects of this mysterious scientific discipline.

A historical characterization

Chapter 2 broadly visualizes the history of psychometrics through a set of academic genealogies, in which the lineages of the presidents of the Psychometric Society are traced back through time. The majority of the presidents descend from influential psychologists like Wilhelm Wundt, William James, James Angell, and Albert Michotte. A smaller yet substantial subgroup of psychometricians stems from mathematician Carl F. Gauss. The genealogies have a strong descriptive purpose and preserve important historical knowledge. However, based on these genealogies we have also formulated a number of interesting aspects about the history of psychometrics. First, a number of very influential individuals, like Francis Galton and Gustav Fechner, are completely missing from the genealogies and others have surprisingly little offspring (like Charles Spearman). This can be for a wide range of reasons, but most of all, it shows that the success of a psychometrician does not have a direct relationship with the number of his or her (mostly his) academic descendants that are involved in the Psychometric Society. Historical importance one the one hand, and a role in the institutionalization of psychometrics as a scientific discipline and of the Psychometric Society on the other, are thus relatively independent factors.

Second, the genealogies expose the fragmentation of psychometrics into a larger branch rooted in psychology, and a smaller branch rooted in mathematics. A suggestion we make here is that the multidisciplinary character of the field is responsible for the tension in psychometrics between psychology- and statistics-oriented researchers, which shows in a number of schisms between the Psychometric Society and other organizations in quantitative psychology. Third, it is shown that over time, psychometrics has become more diverse in terms of socio-cultural background, gender, and substantive background. Here we suggest that the Psychometric Society is now less representative of the contemporary psychometric community than in its early days, when the community was smaller and limited to only a number of universities.

The historical characterization continues briefly in Chapter 6, where we have made a comparison between the use of values in early psychometrics, and the use of values in contemporary psychometrics. Several early psychometricians were explicitly committed to a eugenic ideology and a 'new social order' in which the society was ordered based on ability: people with a high level of intelligence were supposed to end up in demanding positions, e.g. as political leaders, and people with lower levels of intelligence were destined to do less-esteemed manual work. Since intelligence was considered hereditary, the ideal society was a society in which people of high intelligence were encouraged to procreate, and people of low intelligence were discouraged from doing so. Objective measurement of psychological attributes was to become an important method in ranking people based on their ability and assigning them ability-appropriate tasks. A famous example of this is the rise of military testing in WWI, when mental tests were broadly applied to distinguish between prospective soldiers and applicants who were not mentally fit for the job. Psychometrics was thus explicitly led by political and social values. As we show in Chapter 6, contemporary psychometrics is less explicit about its endorsement of values. Since this analysis of values in contemporary psychometrics is more of a philosophical nature, it will be further expanded as part of the philosophical characterization (see below).

An ethnographic characterization

Psychometricians form a group of people who share a culture of research interests, motivations, and practices. Through an oral history study, we aimed to investigate the perspectives of psychometricians on their own field. The purpose of this study was to gain a better understanding of important historical, modern, and possibly future topics that play a role in the field and especially, how psychometricians reflect on these topics. From the interviews, we can conclude that psychometrics is a multifaceted discipline. Some psychometricians are psychology-oriented, in that they find that psychometrics should and could serve psychological theory building. Others are more statistics or data-analytically-oriented, in that they consider psychometrics as a set of statistical models that,

though originally developed for psychological measurement, are essentially transferrable to other research problems and disciplines. Besides the psychologist approach and the data analytical approach, we also found psychometricians who follow a consultant approach, an engineering approach, or a mathematical approach. Psychometricians thus have different attitudes as to what kind of approach should be leading psychometric research.

Besides diversity in terms of psychometric research, we also find that psychometricians reflect diversely on the future of psychometric research. Some psychometricians are optimistic about a future for psychometrics, either because educational testing still fulfills a need in society and psychometricians will thus remain relevant, or because psychometrics has much to offer to and learn from new developments such as data mining, big data, and machine learning. Others are wary of a fruitful future for psychometrics. As a field, it has experienced some serious PR problems. A point for concern for our interviewees is that psychometrics has trouble selling its value both to substantive research and to statistics: psychologists are less inclined to read psychometric literature because of its technical content, while statisticians do not have familiarize themselves with the small research area that is psychometrics. Contemporary psychometrics is thus essentially a pluralist research area that needs to reconnect with related disciplines. Based on the testimonies, we recommend that psychometrics acknowledge its pluralist identity and specify what it aims to achieve in the future.

A philosophical characterization

The third and last characterization is the philosophical characterization, which can be divided in two parts. The first characterization is a causal account of psychometric models, which is given in Chapters 4 and 5. Many psychometric models (including reflective, formative, and network models) imply a certain causal structure which makes these models relevant tools for psychological theory building. These chapters illustrate what it means to give a causal reading of psychometric models and why such a reading can be useful and appropriate in certain contexts. For example, using the common factor model as a causal model is a legitimate choice if the purpose is indeed that of measurement or explanation and when it is assumed that a common causal structure underlies the data. Moreover, a causal interpretation of the factor model legitimizes the focus on *shared*, rather than unique variance of the indicators; and a causal interpretation of the factor model legitimizes the assumption of local independence. The causal account of psychometric models stresses that many of these models imply a certain causal relationship, which can be used for explanatory purposes. However, causal modeling in psychology comes with a number of controversies, such as the causal status of individual differences, interpretational confounding and the problem of generalization (see Chapter 4).

The causal account of psychometric models highlights the explanatory potential of these models: the models can be used to formulate explanations for human behavior. However, this account is not necessarily an accurate description of how psychometric models are being employed in psychometric research. In fact, the realist or causal account of psychometric models as described in these chapters is at odds with the more practical, data-analytical approach that is popular among many contemporary psychometricians.

The second part of the philosophical characterization concerns an analysis of the role of values in contemporary psychometrics (Chapter 6). As mentioned earlier, early psychometrics (late 19th century/early 20th century) was explicit about its political and social motives. Contemporary psychometrics on the other hand, seems a relatively value-free discipline that only deals with issues of a technical nature. However, even the deeply statistical psychometrics of these modern times employs a number of values. For example, it endorses a highly specific notion of fairness, namely that tests or test items are fair when respondents from different groups who have the same measured ability, show the same probabilities for answering an item correctly. Fairness as conceptualized (and 'mathematized') by psychometricians is notably different from other notions of fairness, such as economic or political fairness, and thereby a value that is specific to psychometrics. Other values employed by contemporary psychometrics are the conceptualization of individual differences as quantitative (not qualitative) differences, the aim for objective measurement (it is considered risky to let any personal judgement taint any part of the testing process), and the preference for utility above truth (rather than an interest in the truth-content of psychological theory, psychometricians are often driven by the practical utility of their research).

A portrait of psychometrics

From an outsider's perspective, psychometrics seems quite a mysterious, complicated, yet uniform endeavor. This dissertation aimed to unpack some of the intricacies of this scientific discipline. Together, the three characterizations paint a multidimensional portrait of psychometrics. Psychometrics turns out to be a multidisciplinary and diverse discipline, characterized by a number of methodological approaches and with roots both in psychology and mathematics. This dissertation also shows that psychometrics is a discipline that experiences a number of inner tensions. Psychometrics seems torn between being a subfield of psychology and a discipline that belongs to statistics; between having a social mission and being a mostly technical discipline; between being a field that employs models that are appropriate for hypothesizing about explanations for behavior and a discipline that mostly engages in building useful applications. All in all, I hope this dissertation unlocks some of the complexities of psychometrics, and makes it a more accessible research area for people outside the psychometric community.

Publications

Chapter 2 was published as:

Wijsen, L. D., Borsboom, D., Cabaço, T., Heiser, W. J. (2019). An academic genealogy of Psychometric Society presidents. *Psychometrika, 84,* 562 - 588.

LDW wrote the article, collected the data (the evidence for lineages), revised the article and prepared it for (re)submission. TC helped with collecting data. DB and WJH gave feedback throughout the writing process. All authors helped with the revisions and participated in a thorough fact check.

Chapter 3 is submitted as:

Wijsen, L. D. & Borsboom, D. (under review). Perspectives on psychometrics: Interviews with 20 past Psychometric Society Presidents. Submitted to *Psychometrika.*

LDW and DB deliberated on the study design. LDW wrote the article, collected the data (the interviews), revised the article and prepared it for (re)submission. DB gave feedback throughout the writing process and helped with the revisions.

Chapter 4 is submitted as:

Wijsen, L. D., van Bork, R., Haig, B. D., & Borsboom, D. (in press). *Reflective, formative, and network models: Causal interpretations of three different psychometric models.* (Book chapter in The SAGE Handbook of Theoretical Psychology)

LDW wrote the general framework of the book chapter, including the introduction, some individual paragraphs and the conclusion. RvB, BDH and DB each contributed specific sections. LDW made the final edits and prepared it for submission.

Chapter 5 is published as:

Van Bork, R., **Wijsen, L. D.**, & Rhemtulla, M. (2017). Toward a causal interpretation of the Common Factor Model. *Disputatio, 9*, 581 – 601.

RVB and LDW (both first authors) wrote the draft of the article and prepared the article for submission. MR provided feedback on the manuscript and wrote a number of sections. RVB, LDW, and MR all worked on revisions.

Chapter 6 is submitted as:
Wijsen, L. D., Borsboom, D., & Alexandrova, A. (under review). Values in Psychometrics. Submitted to *Perspectives in Psychometrics*.

LDW wrote the draft of the article and prepared it for submission. AA and DB provided feedback throughout the writing process.

Other publications

Wijsen, L. D. (forthcoming). *Twenty interviews with Psychometric Society presidents: What drives the psychometrician?* Springer.

Borsboom, D. & **Wijsen, L. D.** (2016). Frankenstein's validity monster: The value of keeping science and politics separated. *Assessment in Education: Principles, Policy & Practice, 23,* 281 – 283.

Borsboom, D. & **Wijsen, L. D.** (2017). Psychology's atomic bomb. *Assessment in Education: Principles, Policy & Practice, 24,* 440 – 447.

Short CV

Lisa D. Wijsen was born in Utrecht on January 19th, 1991. Between 2009 and 2013, she obtained a Bachelor's degree in Literature Studies and a Bachelor's degree in Psychology. After that, she obtained a Research Master's degree in Psychology, with a specialization in Psychological Research Methods and a minor in Philosophy of Science. She wrote her master thesis at the University of Canterbury, New Zealand under supervision of Prof. Brian Haig. In June 2015, she received the IOPS Ph.D. grant and in September 2015, she started her Ph.D. project on the history of psychometrics, under supervision of Prof. Denny Borsboom (University of Amsterdam) and Prof. Willem Heiser (Leiden University). Between September 2018 and February 2019, Lisa worked for NWO (the Dutch Research Council) as junior policy officer. During the fourth year of her Ph.D., she visited dr. Anna Alexandrova at the University of Cambridge, which resulted in the paper Values in Psychometrics.

Dankwoord

Dit proefschrift is niet zomaar tot stand gekomen, en ik wil daarom aan een aantal mensen mijn dank betuigen, en allereerst aan mijn zeer enthousiaste en optimistische begeleiders Denny en Willem. Toen wij in September 2015 begonnen aan dit project, hadden we denk ik geen idee waar we eigenlijk aan begonnen: we kenden elkaar niet heel goed, en ook was dit project voor ons alle drie anders dan anders. Maar we hebben het gered! Denny, onze gesprekken waren van onschatbare waarde voor dit proefschrift. Ze gaven me altijd weer nieuwe energie en inspiratie om weer verder te gaan met m'n onderzoek, ook op momenten dat ik vastliep of het hele proefschrift niet meer zag zitten. Je hebt me ook gesteund in die neiging van mij om vooral ook andere dingen dan onderzoek te doen (zoals lesgeven en een uitstapje naar het NWO), en dat jij daar niet van in paniek raakte was voor mij altijd een geruststelling. Ik had schijnbaar al die afleiding eerst nodig om erachter te komen dat ik onderzoek toch wel zag zitten! Daarnaast was het heel fijn om op een afdeling vol met statistische expertise een handlanger te hebben die mij steunde in mijn hang naar een meer filosofische en kwalitatieve aanpak. Willem, ik heb me altijd rijk gerekend met niet één maar twee zeer betrokken begeleiders. Een ontzettende luxe. Jouw oneindige kennis over de psychometrie was een hele belangrijke bron voor mijn proefschrift. Ik heb ook ontzettend fijne herinneringen aan onze afspraken in Leiden en Amsterdam. Al kostte me het in het begin wat moeite om me een weg te banen in die overvloed aan liefde en passie voor de wetenschap, ik kwam er altijd uit met weer nieuwe ideeën en vooral heel veel inspiratie. Mijn dank is groot!

Lieve PML-collega's: ik vind het bijzonder dat zoveel van ons elkaar al kennen sinds onze studententijd en dat we 10 jaar later nog steeds collega's zijn. Ik ben heel blij dat ik nog een tijdje op deze afdeling waar ik me zo thuis voel mag rondlopen, en ik kijk er enorm naar uit om jullie, hopelijk over niet al te lange tijd, weer eens in het echt te zien. Ik mis jullie! Riet, ik heb het altijd heel fijn gevonden om een maatje te hebben op de afdeling, iemand met wie ik op zowel onderwijs- als onderzoeksvlak heel fijn kan samenwerken. Jolanda, of het nou gaat om het aanleren van R-code of haaksteken, je bent altijd enorm behulpzaam en attent, en dat waardeer ik enorm! Ria & Gaby, ik vond het altijd heel erg gezellig in 0.40, en ik vind het heel erg leuk dat we elkaar goed hebben leren kennen en onze ervaringen over het schrijven van een proefschrift kunnen delen. Louise, met jou kon ik het werkelijk over alles hebben (verhuizen naar Deventer, baby's krijgen, champagne drinken) en ik hoop heel erg dat we dat blijven doen. Dylan, vanaf mijn derde Bachelorjaar wist ik dat ik bij jou moest zijn voor een echte peptalk. Alexander, je bent al een tijdje weg, maar de afdeling is simpelweg niet hetzelfde zonder jou. Max, jij was mijn mede-historicus op deze psychologen-afdeling en mijn geesteswetenschappelijke steun

en toeverlaat. Dank voor al die theetjes in Crea! Claudia, ik smacht naar een lekkere lunch in het Bakhuys, laten we het zodra het kan onze lunchtraditie weer oppakken. Heel veel dank aan alle PML'ers en ook aan alle fijne mensen bij psychologie die mijn werk zo leuk maken!

There are a few people who have inspired me throughout my academic career, and since they live all over the world, I'll make a brief transition to English. Anna, my time at Cambridge University turned out to be one of the most enjoyable and inspiring times of my life. I came back to Amsterdam with a renewed love for the academic live, due to your incredibly lively department and our pleasant collaboration. I'm really pleased we still get to collaborate on our paper, and I'm hoping to continue our collaboration in the future. Brian, I was so incredibly lucky to write my Master thesis in New Zealand with you as my supervisor. It was a total gamble; stepping on a plane, flying to the other side of the world, hoping that it was going to be a positive experience, but it certainly was. You were a great supervisor, and I'm really pleased that we are still in touch. Dear presidents of the Psychometric Society, your academic lineages and perspectives on the field of psychometrics form an integral part of this dissertation. I want to thank you for responding so kindly to all of my requests, to participating in my interview project, and to all of your editing efforts over the last couple of years. I'm looking forward to sharing my research with you! Also, special thanks to IOPS and the Psychometric Society for funding my research.

Lieve Denise, Laura en Sharon: jullie zijn mijn Coffee Company maatjes (ooit zal ik ook eens een koffietje uitproberen tijdens een bezoekje aan de CC). Bij jullie kan ik terecht wanneer mijn motivatie naar het nulpunt is gezakt of om gewoon even lekker bij te kletsen, uiteraard met een drankje erbij. Ik kijk enorm uit naar het moment dat wij onze koffiemomenten en borrels met kaasplankjes, port en garnalenkroketjes weer op kunnen pakken! Lieve Crispijn, Charlotte en Jasmijn: we hebben elkaar tijdens de intreeweek ontmoet, en de jaren die daarop volgden waren absoluut een hoogtepunt van mijn leven! Het doet nog steeds pijn dat ik er niet bij was toen ons intreeweekgroepje zich met z'n tienen in een auto probeerde te proppen. Ik vind het heel leuk dat we nog steeds zulk goed contact hebben en ik kan niet wachten op een ouderwetse wijn- of spelletjesavond.

Renske & Jessica, ik ben ontzettend blij dat jullie mijn paranimfen zijn. Lieve Renske, toen we elkaar voor het eerst ontmoetten wisten we eigenlijk niet wat we met de ander aan moesten en hebben we gevochten om de gunsten van Crispijn, maar een proefschrift schrijven is toch wel iets wonderlijks: opeens hadden we iets gemeen en ontstond onze vriendschap. Gelukkig hebben we besloten om Crispijn dan maar gewoon te delen. We

zijn ontzettend anders maar hebben daarom misschien juist ook heel veel aan elkaar. Lieve Jessica, het is soms een gek idee om te bedenken dat we pas jaren nadat we allebei van dezelfde school afkwamen bevriend werden, beginnend met een borrel in de legendarische Brabantse Aap. Onze studies in de letteren bleken genoeg voor een eerste binding en die is sindsdien gebleven. Het is ontzettend fijn om met jou m'n passie voor lezen te kunnen delen, als wel al het andere dat op ons pad komt in onze levens. Ik hoop dat ik nog heel lang met jullie bevriend mag blijven!

Lieve Wim. Ruim drie jaar geleden ben ik tegen jou opgebotst in de Bagels & Beans, en inmiddels ben je niet meer weg te denken uit mijn leven. Inmiddels wonen we samen (op 500 meter afstand van deze Bagels & Beans), en zijn we volledig geïntegreerd in het IJburgse leven, zeker nu een kleine Pommelien of Pommelino op komst is. De afgelopen drie jaar waren fantastisch en ik heb ontzettend veel zin in ons leven samen; ik hoop – ik weet – dat er nog heel veel moois voor ons ligt.

Lieve Sophie. Wij groeiden op als twee zusjes die totaal verschillend waren, en wie van ons nou precies geadopteerd was, was wel eens onderwerp van gesprek. Ik ben zo blij dat we naarmate we ouder zijn geworden we ook steeds hechter zijn geworden, en dat ik zo'n fijne relatie heb met mijn zus, waarvan ik weet dat ze altijd voor mij in de bres zal springen wanneer iemand het ook maar waagt om mij onrecht aan te doen.

Als laatste wil ik heel graag mijn ouders bedanken. Ik heb enorm veel geluk gehad in mijn leven om op te groeien in zo'n fijn gezin als dat van ons. Inmiddels is de samenstelling van ons gezin wel wat veranderd, maar wat nooit zal veranderen is hoe betrokken jullie nog steeds zijn in mijn leven en in dat van Sophie. Jullie hebben me alle kansen en liefde gegeven, en me altijd onvoorwaardelijk gesteund in al mijn keuzes. Ik gun iedereen zo'n fijne jeugd en ouders als die van mij.