Structural Equation Models: From Paths to Networks

2nd Edition

J. Christopher Westland

2018-12-26

Contents

Foreword the Second Edition of Structural Equation Models

Since publication of the first edition of Structural Equation Models, I have been fortunate to maintain an active dialog on structural equations modeling (SEM) with many of my colleagues around the world. I never cease to be surprised with the broad divergence of opinions and myriad applications of SEM methodologies. Statistical methods such as regression and ANOVA rely on datasets of objectively measured constructs. These fail to satisfy a widespread need among researchers to analyze data concerning relationships between hypothesized but unobservable constructs: aesthetics, perceptions, utilities and other human and social constructs. Criticisms of SEM arisen from its lack of fit statistics and indeed the lack of defensible sampling strategies. But my argument is that these problems can be repaired while retaining the desirable features of SEM.

SEM has been applied in both the natural and the social sciences, but it has proven particularly valuable in the social sciences, where researchers apply SEM approaches rather than more structured regression approaches by the inclusion of unobservable (or latent) constructs and by the use of computationally intensive iterative searches for coefficients that fit the data. The expansion of statistical analysis to encompass unmeasurable constructs using SEM, canonical correlation, Likert scale quantification, principal components and factor analysis have vastly extended the scope and relevance of the social sciences over the past century. Subjects that were previously the realm of abstract argumentation have been transported into the mainstream of scientific research (see I. Allen and Seaman 2007; Altman and Royston 2000).

This new edition of this widely cited book surveys the full range of available structural equation modeling (SEM) methodologies. The book has been updated throughout to reflect the arrival of new software packages, which have made analysis much easier than in the past. Applications in a broad range of disciplines are discussed, particularly in the social sciences where many key concepts are not directly observable. This is the first book to present SEM's development in its proper historical context–essential to understanding the application, strengths and weaknesses of each particular method. This book also surveys emerging approaches that complement SEM. They have been applied in diverse areas in engineering, including neuroscience for accurate examination of the activity among neural regions during different behaviors. The partial least squares SEM method was contemporaneously developed with PLS regression to address problems in chemistry and spectrography. They improve on predecessor path models that were widely used in genetic research in livestock and agriculture and environmental studies in the elicitation of ecological networks. SEM's ability to accommodate unobservable theory constructs through latent variables is of significant importance to social scientists. Latent variable theory and application are comprehensively explained and methods are presented for extending their power, including guidelines for data preparation, sample size calculation and the special treatment of Likert scale data. Tables of software, methodologies and fit statistics provide a concise reference for any research program, helping assure that its conclusions are defensible and publishable.

J. Christopher Westland
Chicago 2018-12-26

Chapter 1

An Introduction to Structural Equation Models

The past two decades have witnessed a remarkable acceleration of interest in structural equations modeling (SEM) methods in many areas of research. In the social sciences, researchers often distinguish SEM approaches from more powerful systems of regression equation approaches by the inclusion of unobservable constructs (called latent variables in the SEM vernacular), and by the use of computationally intensive iterative searches for coefficients that fit the data. The expansion of statistical analysis to encompass unmeasurable constructs using SEM, canonical correlation, Likert scale quantification, principal components and factor analysis have vastly extended the scope and relevance of the social sciences over the past century. Subjects that were previously the realm of abstract argumentation have been transported into the mainstream of scientific research (see I. Allen and Seaman 2007; Altman and Royston 2000).

Statistical methods to identify latent constructs underlying observations evolved in the 1930s. Principal component analysis (PCA), factor analysis and other methods look for methods to reduce the dimensionality of a complex multicolinear dataset. Latent factors accounting for most of the similarity or distance of measurements could potentially be inferred from these factors. SEM methods grew out of efforts to infer additional structure between these latent constructs.

Many of the seminal studies on structural statistical models in economics took place in the Cowles Commission (then at the University of Chicago) in the 1940s and 1950s and later in the Chicago school of economics from the 1950s on. In a 1976 paper, Robert Lucas of the Chicago school argued that generic additive linear models such as those invoked in the panel regressions commonly used in econometrics, lacked stability and robustness (Lucas Jr 1992). He argued, in what has come to be known as the 'Lucas critique', that empirical models are improved when constructs are policy-invariant i.e. structural, implying that they would be unlikely to change whenever the competitive environment or a particular policy changed. Lucas suggested that researchers need to model the "deep structural parameters" (relating to preferences, technology, and resource constraints) that are assumed to govern individual behavior. Structural models in (Lucas Jr 1992) were intended to enable a positive research program for econometrics, allowing for prediction and real-world decisions. Policy-invariant structural models are constructed through analysis of the underlying dynamics of the construct relationships and behavior, and are based on a 'theory' of how the real-world works. The 'Lucas critique' promoted *a priori* theory building, and this has become common practice in structural equation modeling. It is now standard practice to design the theorized causal structures in an SEM, whether the statistical method is PLS-PA, LISREL or regression approach, prior to statistical estimation.

The products of SEM statistical analysis algorithms fall into three groups:

1. pairwise canonical correlations between pairs of prespecified latent variables computed from observable data (from the so-called partial least squares path analysis, or PLS-PA approaches);
2. multivariate canonical correlation matrices for prespecified networks of latent variables computed from observable data (from a group of computer intensive search algorithms originating with Karl Jöreskog); and
3. systems of regression approaches that fit data to networks of observable variables whose clusters are hypothesized to co-vary with latent constructs. Other methods of latent variable analysis are now emerging with the introduction of machine learning new social network analysis.

Many of the PLS-PA algorithms are variations on an incompletely documented software package described in (Lohmoller 1988; Lohmöller 1989; Lydtin et al. 1980) and we sometimes still see some of their old Fortran code inside a customized user interface wrapper. Fortunately Monecke and Leisch (2012a) has incorporated Wold's mathematics in their thoroughly modern semPLS package for R. PLS-PA has a tendency to be confused with Wold's partial least squares regression – a problem Herman Wold tried unsuccessfully to correct. The path analysis PLS-PA commonly used in latent variable investigations is unrelated to Wold's (Wold 1966; Hill 1979) partial least squares regression methods, instead being a variation on Wold's (Wold 1966; Hill 1979) canonical correlation methods to elicit the correlations of latent variables.
Two different covariance structure algorithms are widely used: (1) LISREL (an acronym for LInear Structural RELations) (Joreskog and Van Thillo 1972; Jöreskog 1993; Jöreskog and Sörbom 1982; Joreskog, Sorbom, and Magidson 1979; and Joreskog 1970) and the AMOS (Analysis of Moment Structures) (Fox 2006; McArdle and Epstein 1987; McArdle 1988) Variations on these algorithms have been implemented in EQS, TETRAD and other packages.

Methods in systems of equations modeling and social network analytics are not as familiar in the social sciences as the first two methods, but offer comparatively more analytical power. Accessible and comprehensive tools for these additional approaches are covered in this book, as are research approaches to take advantage of the additional explanatory power that these approaches offer to social science research.

The breadth of application of SEM methods has been expanding, with SEM increasingly applied to exploratory, confirmatory and predictive analysis through a variety of ad hoc topics and models. SEM is particularly useful in the social sciences where many if not most key concepts are not directly observable, and models that inherently estimate latent variables are desirable. Because many key concepts in the social sciences are inherently Two different covariance structure algorithms are widely used: (1) LISREL (an acronym for LInear Structural RELations) (Joreskog and Van Thillo 1972; Jöreskog 1993; Jöreskog and Sörbom 1982; Joreskog, Sorbom, and Magidson 1979; and Joreskog 1970) and the AMOS (Analysis of Moment Structures) (Fox 2006; McArdle and Epstein 1987; McArdle 1988) Variations on these algorithms have been implemented in EQS, TETRAD and other packages.

Methods in systems of equations modeling and social network analytics are not as familiar in the social sciences as the first two methods, but offer comparatively more analytical power. Accessible and comprehensive tools for these additional approaches are covered in this book, as are research approaches to take advantage of the additional explanatory power that these approaches offer to social science research.

The breadth of application of SEM methods has been expanding, with SEM increasingly applied to exploratory, confirmatory and predictive analysis through a variety of ad hoc topics and models. SEM is particularly useful in the social sciences where many if not most key concepts are not directly observable, and models that inherently estimate latent variables are desirable. Because many key concepts in the social sciences are inherently

1.1 Latent Constructs as Organizing Principles of Science in the 20th Century

In science, an idea is a hypothesis that gives structure to our observations. Ideas are latent constructs embellished with mechanisms to test, use, predict and control their implementation. Three ideas revolutionized science in the 20th century: the atom, the bit and the gene.

The atom provided an organizing principle for 20th century physics. Hypotheses about the atom date from the Greek philosopher Democritus, and steady advancements marshaled the evolution of chemistry out of alchemy. But it was Einstein's obsession with determining the size of an atom that indirectly motivated his 1905 "annus mirabilis" when he published groundbreaking papers on the photoelectric effect, Brownian motion and special relativity and the equivalence of mass and energy, seminal works in 20th century physics and chemistry.

The gene is innately human. Its origins have seduced the attention of philosophers and politicians, more often than not, leading them astray. Genes are the unseen first cause of human and animal 'phenotypes' – their observable, externalized consequences resulting from interaction of an organism's genotype with the environment. Phenotypes manifest themselves as morphology, skin color, strength and numerous other characteristics. The word 'gene' was coined by botanist Wilhelm Johannsen as a shortening of Darwin's pangene.

The search for the unobservable genes that would lead to various desirable or undesirable phenotypes has been a major factor in the history and philosophy of mankind. In the 21st century, the quest to master genetics has enlisted our knowledge of atoms and bits as well. The physicist John Wheeler famously stated that "...all things physical are information-theoretic in origin", a sentiment that drives much of modern genetics, and brings us to the last of our 20th century 'ideas'.

The bit, a portmanteau of "binary digit" arose from efforts to quantify and encode information, particularly in such devices as the Jacquard looms in the early 1800s. Attempts to improve bandwidth in telegraph lines in the mid-19th century led to speculation that there existed some sort of fundamental measure of information: a bit. Bits were fundamental to Morse code, and the basis for Hartley's and Shannon's seminal work on information theory.

It was not originally the desire to make better men or women that spurred developments in the science of genetics; it was mans desire to improve domesticated crops and animals.

1.2 Path Analysis in Genetics

Though structural equation models today are usually associated with soft problems in the social sciences, they had their origin in the natural sciences – specifically biology. Europe's 19th century scholars were challenged to make sense of the diverse morphologies observed during an age of explorations, in Asia, Africa and the Americas, as well as at home. In this period, new species of plants and animals were transplanted, domesticated, eaten and bred at an unprecedented rate. An American ultimately provided one statistical tool that allowed scholars to build a science out of their diverse observations.

Seldom has a non-human animal been so thoroughly poked, observed, trained and dissected as the domesticated dog. A member of the Canidae family, the dog is distantly related to coyotes and jackals, dingoes, foxes and wolves. There is evidence of distinct dog breeds as early as five thousand years ago in drawings from ancient Egypt. The business of designing dogs for particular purposes began in earnest around the sixteenth century, and by the nineteenth century, clubs and competitions abounded for the naming and monitoring of breeds. There is a huge variation of sizes, shapes, temperaments and abilities in modern dogs – much more so that in their homogeneous wolf ancestors. This has resulted from humans consciously influencing the genetics of dog populations through an involved network of interbreeding and active selection. But none of this was a science at the dawn of the 20th century, despite enormous expenditures, and centuries of breeding and contests to create 'the perfect dog.' There was no theory (or perhaps too many competing but unsupported theories) about how particular characteristics arose in a particular sub-population of dogs. The sciences of evolution and genetics seldom spoke to each other before the 20th century. The most influential biologists held the idea of blending inheritance, promoted in a particular form in Charles Darwin's theory of pangenesis – inheritance of tiny heredity particles called gemmules that could be transmitted from parent to offspring. In those days, the work of the Augustinian friar and polymath Gregor Mendel was unknown, having been rejected and forgotten in the biology community when published in the 1860s. Mendel's sin was to introduce mathematics into a field that biologists felt should be a descriptive science, not an analytical one. Rediscovery of Mendel's writings in the early 20th century led biologists towards the establishment of genetics as a science and basis for evolution and breeding.

Geneticist, Sewall Wright, along with statisticians R. A. Fisher and J.B.S. Haldane, were responsible for the modern synthesis that brought genetics and evolution together. Wright's work brought quantitative genetics into animal and plant breeding, initiating the hybrid seed revolution that transformed US agriculture in the first half of the 20th century. Wright actively mapped the breeding networks that created desirable hybrids – of particular significance to the dog breeders was Wright discovery of the inbreeding coefficient and of methods of computing it in pedigrees. The synthesis of statistical genetics into the evolution of populations required a new quantitative science with which to map the networks of influence, on random genetic drift, mutation, migration, selection, and so forth. Wright's quantitative study of influence networks evolved in the period 1918 through 1921 into Wright's statistical method of path analysis – one of the first statistical methods using a graphical model, and one which is the subject of this book. Let's begin by reviewing the evolution of path analysis from the dark ages of 19th century evolution debates, through today's statistical methods, to emerging techniques for mapping the extensive networks of biological interactions important to genetics and biotechnology in the future.

1.3 Sewall Wright's Path Analysis

Path analysis was developed in 1918 by geneticist Sewall Wright (Wright, 1920, 1921, 1934) who used it to analyze the genetic makeup of offspring of laboratory animals.

Early graphs were very descriptive, with pictures and stories attached. But gradually pictures of laboratory critters gave way to representative boxes and positive or negative correlations

Rensis Likert's work at the University of Michigan in the 1930s and 1940s saw path analysis directed towards social science research. Social scientists need to model many abstract and unobservable constructs – things like future intentions, happiness, customer satisfaction, and so forth. Though not directly observable, there typically exist numerous surrogates that can provide insight into such abstract (or latent) constructs – these observable surrogates are called 'indicators' of the latent variable. Further innovation in path models evolved around Hermann Wold's extensions of Hotelling's seminal work in principal components analysis (PCA). Wold began promoting the principal components as representations of abstract (latent) constructs. Latent abstractions proved useful in the evolving fields of psychometrics and sociological surveys, and were widely adopted in the 1950s and 1960s (Hotelling, 1936; Wold, 1966). Path diagrams evolved once again, to incorporate Wold's conceptualization of latent constructs as the first component from a principal components analysis. Wold called the network model of latent variables the 'structural model' or sometimes the 'inner' model. The term 'structural equation model' came about from his use, which Wold borrowed from the matrix terminology of systems of equation regression approaches developed at the Cowles Commission. Social scientists were ultimately not content to let PCA dictate their choice of abstractions. In education research, Henry Kaiser and Lee Cronbach, both faculty in the University of Illinois, School of Education in the 1950s argued that such abstract concepts could be conceived prior to data collection, and the collected data with the abstract concept could be reviewed after the fact to see that it actually looks like a first principal

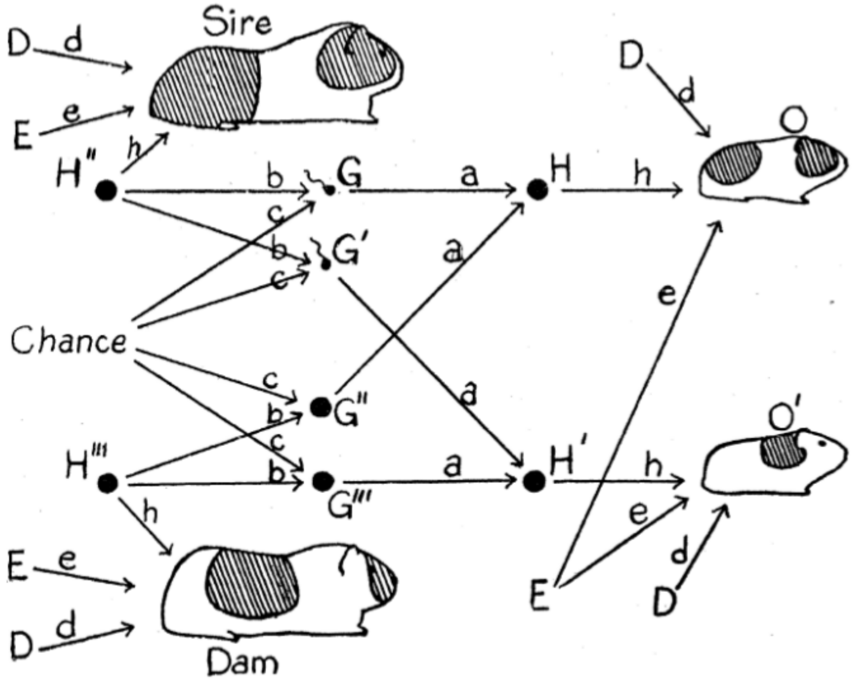

Figure 1.1: Relations between litter mates and their parents (H represents (latent) genetic components, other capital letters are (manifest) environmental factors, and lower case letters are path coefficients

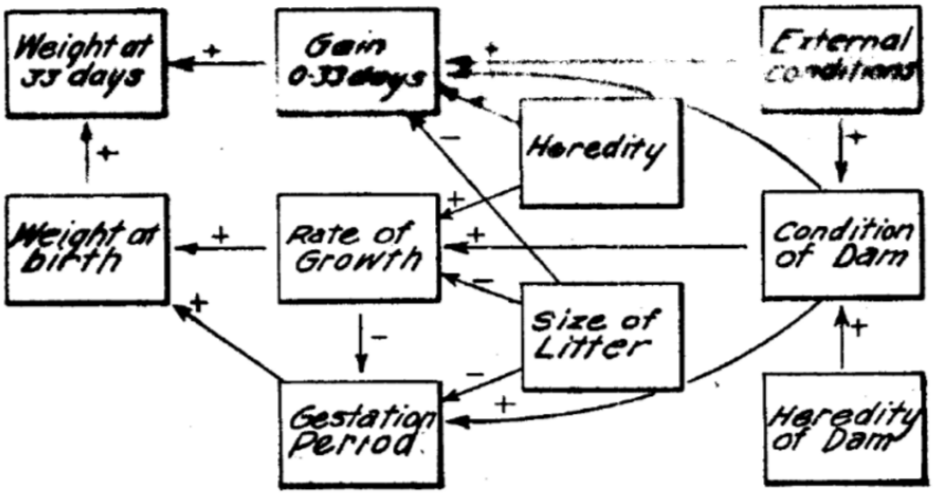

Figure 1.2: Interrelationship between the factors that determine weight of guinea pigs at birth and at weaning

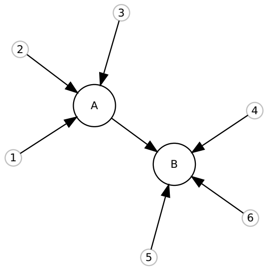

Figure 1.3: (#fig:modern_diagram)A generic path model with latent variables

component. These alternative approaches to defining the relationship between indicators and the latent variables they indicate created what Wold called formative and reflective links. If the researcher chooses the indicators before the latent variables, the links are called 'formative' because the factor (link) weights and the latent variable are formed from the first component of the PCA. If the researcher chooses the latent construct before the indicators, the links are called 'reflective' because the factor (link) weights are believed to reflect the abstract (latent) construct. By the 1960s ad hoc path diagrams had evolved to neat boxes and bubbles that identified measurable and latent components of the model

1.4 Networks and Cycles

Many real world influences are cyclic, and interesting questions revolve around the convergence to equilibrium – e.g., in predator-prey ratios, and corporate 'satisficing' and so forth. Path tracing has become an essential feature of the graphical interface for SEM software programs. These specific rules are designed to yield graphs (and thus models) that are non-recursive – i.e., do not have influence loops. Consider for example figure 1.4, a graph of three variables A, B, and C with a recursive relationship.

Assume that the correlation between each pair of latent variables in the figure is 0.5; thus a change in one variable results in a linear influence of 50% of that change on the variable that the outgoing arrow points to. Then we might ask 'what will be the net effect on all variables (including A) if we vary A by 1.0?' The variance of A will be affected by the initial perturbation; 50% of that will appear at B, 25% at C and 12.5% at A again, and so forth. This is not a result that can be teased out with regressions (nor with PLS path analysis).

The expected correlation due to each chain traced between two variables is the product of the standardized path coefficients, and the total expected correlation between two variables is the sum of these contributing path-chains. Intrinsically, Wright's rules assume a model without feedback loops it puts paid to the mental simplification of a simple linear sequence of causal pathways. Path modeling software will generally not allow the design of network graphs with cycles – the graphs will have to be acyclic. Wright's rules are designed to insure this. In order to validly calculate the relationship between any two boxes proposed a simple set of path tracing rules for calculating the correlation between two variables (Wright 1934). The correlation is equal to the sum of the contribution of all the pathways through which the two

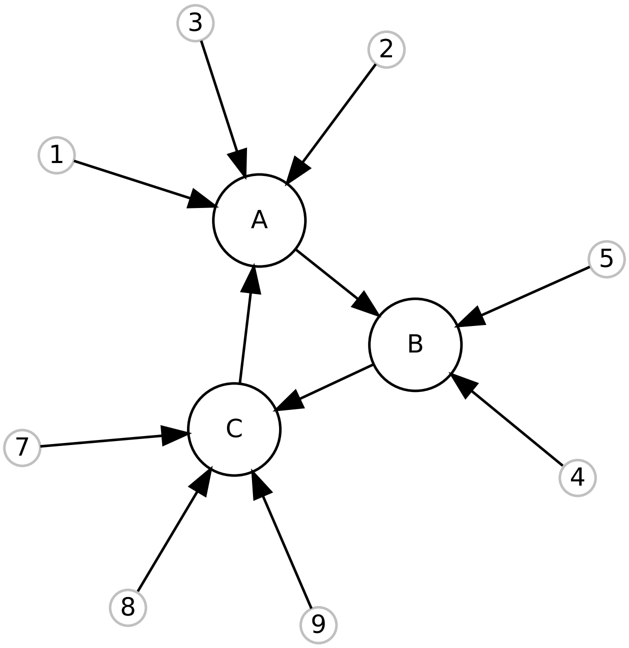

Figure 1.4: A cyclic or recursive graph

variables are connected. The strength of each of these contributing pathways is calculated as the product of the path-coefficients along that pathway. The rules for path tracing are based on the principle of never allowing flow to pass out of one arrow head and into another arrowhead. These are:

1. You can trace backward up an arrow and then forward along the next, or forwards from one variable to the other, but never forward and then back.
2. You can pass through each variable only once in a given chain of paths.
3. No more than one bi-directional arrow can be included in each path-chain.

These three rules assured Wright that he would estimate paths on a directed acyclic graph. That is, a network model formed by a collection of vertices and directed edges, each edge connecting one vertex to another, such that there is no way to start at some vertex and follow a sequence of edges that eventually loops back to that vertex again. Directed acyclic graphs pose significantly fewer problems to mathematical analysis, and restricting path analysis allows simpler calculations, with little loss of generality. Introduction of computer intensive analysis of variance approaches by (Joreskog 1970) allowed more general network models ultimately to be estimated.

1.5 What is a Path Coefficient?

Wright was satisfied estimating his paths with correlation coefficients. This made a great deal of sense before the era of modern computers (or even hand calculators). Furthermore, he argued that in many cases, any other path coefficients required could be recovered from transformations of these correlation coefficients (Wright 1960). Nonetheless, as path modeling grew more common, statisticians began experimentation with covariance and regression coefficients. This evolution started with a suggestion in the early 1950s. Tukey (see (Cochran, Mosteller, and Tukey 1954)) advocated systematic replacement in path analysis of the dimensionless path coefficients by the corresponding concrete path regressions. Geneticists (Turner and Stevens 1959) published the seminal paper presenting the mathematics of path regression and inviting extensions that would apply other developments in statistics. In the early days of data processing, both Hermann Wold and his student Karl Jöreskog developed software adopting variations of Turner and Stevens' mathematics that took advantage of computer intensive techniques developed in the 1970s and 1980s. Over time, the candidate list for path coefficients has steadily broadened. Here is a short list of some of the alternatives that are popular:

1. Pearsonean correlation (Wright 1921)
2. Canonical correlation between latent variables (Joreskog 1970)(Joreskog and Van Thillo 1972)(Wold 1966)(Wold 1974)
3. Regression coefficients (Hill 1992)
4. Covariance (Joreskog and Van Thillo 1972)
5. Systems of equation regression coefficient (TW Anderson 2005a)(Anderson and Gerbing 1988)(A. Zellner and Theil 1962; Zellner 1962)
6. Generalized distance measures (Kailath 1967)

The work of Tukey and Turner and Stevens along with the burgeoning availability of scientific computers in universities ushered in an era of ferment, where computationally intensive methods were invoked to develop ever more complex (and possibly informative) path coefficients.

1.6 Applications and Evolution

Geneticist Sewall Wright was extremely influential, and in great demand as a reviewer and editor. His work was widely cited and served as a basis for a generation of statistical analysis in genetic and population studies in the natural sciences. Wright's work influenced early 1940s and 1950s Cowles' Commission work on simultaneous equations estimation centered on Koopmans' algorithms from the economics of transportation and optimal routing. This period witnessed the development of many of the statistical techniques we depend on today to estimate complex networks of relationships between both measured (called factors or exogenous, manifest or indicator variables) and imagined variables (called latent or endogenous variables). After Sewall Wright's seminal work, structural equation models evolved in three different streams: (1) systems of equation regression methods developed mainly at the Cowles Commission; (2) iterative maximum likelihood algorithms for path

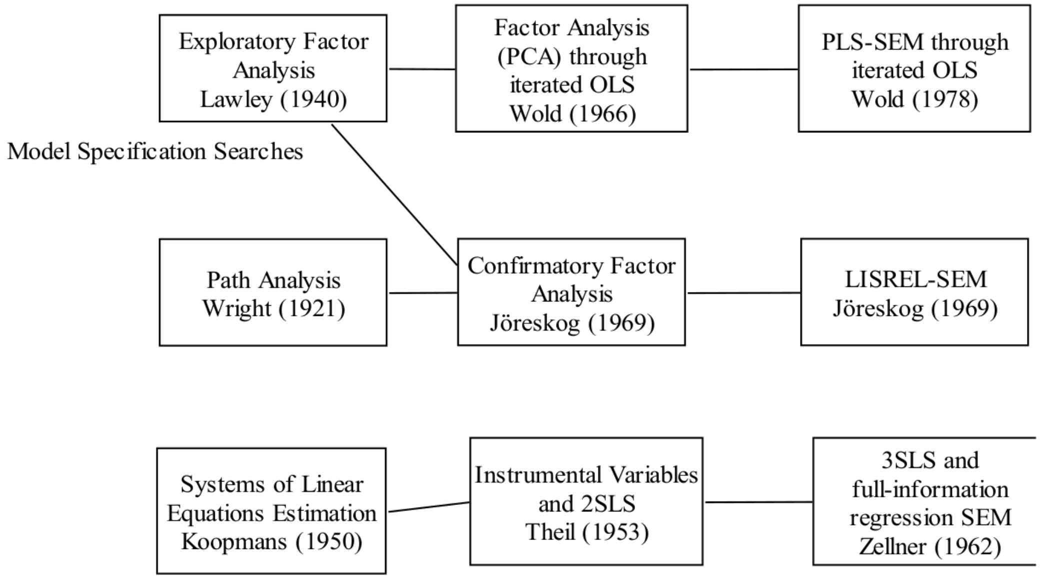
Exploratory Factor Analysis Lawley (1940)
Factor Analysis (PCA) through iterated OLS Wold (1966)
PLS-SEM through iterated OLS Wold (1978)
Model Specification Searches
Path Analysis Wright (1921)
Confirmatory Factor Analysis Jöreskog (1969)
LISREL-SEM Jöreskog (1969)
Systems of Linear Equations Estimation Koopmans (1950)
Instrumental Variables and 2SLS Theil (1953)
3SLS and full-information regression SEM Zellner (1962)

Figure 1.5: History of Structural Equation Models and thier precursors

analysis developed mainly at the University of Uppsala; and (3) iterative least squares fit algorithms for path analysis also developed at the University of Uppsala. Figure 1 describes research book dates of the pivotal developments in latent variable statistics in terms of method (pre-computer, computer intensive and a priori SEM) and objectives (exploratory / prediction or confirmation).

Figure 1.5 describes the relation between SEM methods in terms of how they use data (limited vs. full information methods) and specific methods, grouped by the solutions they converge to under relatively weak regularity assumptions on the data. Both LISREL (an acronym for linear structural relations) and partial least squares (PLS) were conceived as computer implementations, with an emphasis from the start on creating an accessible graphical and data entry interface – and extension of Sewell Wright's path analysis diagrams – to ensure widespread usage. Additionally they were designed to incorporate 'latent factors' following the example of three other multivariate alternatives which involved latent constructs: (1) discriminant analysis – the prediction of group membership from the levels of continuous predictor variables; (2) principal components regression – the prediction of responses on the dependent variables from factors underlying the levels of the predictor variables; and (3) canonical correlation – the prediction of factors underlying responses on the dependent variables from factors underlying the levels of the predictor variables. In an interesting historical footnote, the application of computers to Wold's partial least squares regression (PLSR) version of path analysis was short lived. PLSR computes regression coefficients where data are highly multicolinear (and finds modern use in spectrographic analysis). But this is seldom a problem in path analysis, and when Wold's research associate Lohmöller released his desktop computing software LVPLS that implemented both Jöreskog's as well as his own interpretation of Wold's ideas for path modeling, he implemented ordinary least squares for computing path coefficients, because it was faster to computer, and yielded the same coefficient values. Lohmoller's algorithms (Lohmöller 1989)(Lohmöller 1989) are used in modern PLS path analysis packages and even though PLS path analysis includes PLS in the name, it does not use PLSR in its algorithm. Wold actually tried to change the acronym PLS, but the name stuck (Wold 1974).

1.7 The Chicago School

Regression methods date back to the early 19th century, when Legendre and Gauss developed least squares regression, correlation methods, eigenvalues, eigenvectors, determinants, and matrix decomposition methods. All were extensions of Gauss' theory of algebraic invariants, used initially to fit data on orbits of astronomical objects, where much was known about the nature of errors in the measurements and in the equations, and where there was ample opportunity for comparing predictions to reality. Astronomer and social scientist Adolphe Quetelet was among the first who attempted to apply the new tools to social science, planning what he called a 'social physics', a discipline that was well evolved by the late 19th century. The resulting tools provided the basis for what would eventually become the study of econometrics. Alfred Cowles made his fortune in the insurance industry in the 1920s, and managed to keep that fortune intact after 1929. During a lengthy hospital stay, he started collecting financial data (difficult in the days before the SEC and auditing), a project which ultimately grew into the modern Compustat / CRSP databases. In 1932 he used his fortune to set up the Cowles Commission (first at Cowles' alma mater Yale, later moving to Chicago when Nobelist Tjalling Koopmans moved, and later back to Yale) to develop econometric statistical models for his databases. From his own experience in investment counseling, he had been frustrated by the 'guessing game' techniques used in forecasting the stock market and believed that scholarly research by experts could produce a better method, and suggested as much in a famous book 'Can Stock Market Forecasters Forecast?' (Cowles 1933)(J. Westland 2010). Early Cowles' work on simultaneous equations estimation centered on Tjalling Koopman's algorithms from the economics of transportation and optimal routing. SEM work centered on maximum likelihood estimation, and closed form algebraic calculations, as iterative solution search techniques were limited in the days before computers. (Anderson and Rubin 1949)(Anderson and Rubin 1950) developed the limited information maximum likelihood (LIML) estimator for the parameters of a single structural equation, which indirectly included the 2SLS estimator and its asymptotic distribution (TW Anderson 2005a)(Anderson, Kunitomo, and Matsushita 2010)(Farebrother 1999). Two-stage least squares (2SLS) was originally proposed as a method of estimating the parameters of a single structural equation in a system of linear simultaneous equations, being introduced by (Theil 1980) (Theil and Boot 1962) and more or less independently by (Basmann 1957) and (Sargan 1958). Anderson's LIML was eventually implemented in a computer search algorithm, where it competed with other iterative SEM algorithms. Two stage least squares regression (2SLS) was by far the most widely used systems of equations method in the 1960s and the early 1970s. The explanation involves both the state of statistical knowledge among applied econometricians and the primitive computer technology available at the time. The classic treatment of maximum likelihood methods of estimation is presented in two Cowles Commission monographs: (Turner and Stevens 1959) and (Koopmans 1951). By the end of the 1950s computer programs for ordinary least squares were available. These programs were simpler to use and much less costly to run than the programs for calculating LIML estimates or other approaches requiring iterative search for solutions. Owing to advances in computer technology, and, perhaps, also the statistical background of applied econometricians, the popularity of 2SLS started to wane towards the end of the 1970s. Computing advances meant that the difficulty of calculating LIML estimates was no longer a daunting constraint. Zellner (A. Zellner and Theil 1962; Zellner 1962) algebraically extended 2SLS to a full information method using his seemingly unrelated regressions technique, calling the result three stage least squares (3SLS). In general, the Chicago approaches were couched as systems of equations, without the familiar 'arrow and bubble' diagrams we have come to associate with SEM. Geneticist Sewall Wright conducted a seminar in the 1940s on path coefficients to the Cowles Commission emphasizing the graphical 'bubble and arrow' diagrams that he used in path analysis, and which have since become synonymous with path analysis and its extensions to SEM. Unfortunately, neither Wright nor the Cowles econometricians saw much merit in the other's methods and the main body of Cowles research continued to be dominated by the algebra of Tjalling Koopmans' systems of equations and optimal routing perspective (Christ 1994).

1.8 The Scandinavian School

Structural equation modeling (SEM) statistical methods provide statistical fitting to data of causal models consisting of unobserved variables. SEM approaches were computer intensive attempts to generalize Sewall Wright's path analysis methods, motivated by the merging of path analysis with the systems of equations econometrics of the Cowles Commission. Sewall Wright was also involved in applications since the 1920s of factor analysis – the clustering of observations around theorized but unobserved factors – was accomplished through various ad hoc procedures. Statistical procedures for factor analysis were formulated in (Hotelling 1933) and (Thurstone 1935) who proposed algorithms for maximum likelihood (ML) factor analysis. The computationally intensive iterative solution of Lawley's algorithms needed to wait for computing power that only became available at the end of the 1960s (Kaiser 1960). Path analysis was applied much later in sociology and psychology (see (Werts, Linn, and Jöreskog 1974)(Werts, Jöreskog, and Linn 1972)). The latter two studies introduced many of the new ideas from econometrics to areas where 'latent' factors played a major part in theory. Ideas congealed gradually between the mid-1960s and the mid-1980s when most of the vocabulary we would recognize today was in place. SEM path modeling approaches currently in vogue were largely developed at the Cowles Commission building on the ideas of the geneticist Wright, and championed at Cowles by Nobelist Trygve Haavelmo. Unfortunately, SEM's underlying assumptions were challenged by economists such as (Freedman 1987) who objected that SEM's 'failure to distinguish among causal assumptions, statistical implications, and policy claims has been one of the main reasons for the suspicion and confusion surrounding quantitative methods in the social sciences.' Haavelmo's path analysis never gained a large following among U.S. econometricians, but was successful in influencing a generation of Haavelmo's fellow Scandinavian statisticians. Hermann Wold (University of Uppsala; Wold completed his doctoral thesis under the renowned statistician Harold Cramer) developed his Fixed-Point and PLS approaches to path modeling; and Karl Jöreskog (University of Uppsala) in the development of LISREL maximum likelihood approaches. Both methods were widely promoted in the US by University of Michigan marketing professor Claes Fornell (Ph.D., University of Lund) and his firm CFI, which has conducted numerous studies of consumer behavior using SEM statistical approaches. Fornell introduced SEM techniques to many of his Michigan colleagues through influential books with David Larker (Fornell and Larker 1981) in accounting, Wynne Chin and Don Barclay in information systems (Barclay, Higgins, and Thompson 1995; Chin 1998; Chin and Newsted 1999), Richard Bagozzi (Bagozzi and Warshaw 1990) in marketing and Fred Davis for validating the technology acceptance model (Davis, Bagozzi, and Warshaw 1992), (Davis, Bagozzi, and Warshaw 1989). Development of SEM statistics in Sweden occurred in the late 1960s when computing power was just becoming widely available to academics, making possible computer intensive approaches to parameter estimation. (Jöreskog 1993; Joreskog 1970; Joreskog, Sorbom, and Magidson 1979; Joreskog and Van Thillo 1972) developed rapidly converging iterative methods for exploratory ML factor analysis (i.e., the factors are not defined in advance, but are discovered by exploring the solution space for the factors that explain the most variance) based on the Davidon-Fletcher-Powell math programming procedure commonly used in the solution of unconstrained nonlinear programs. As computing power evolved, other algorithms became feasible for searching the SEM solution space, and current software tends to use Gauss-Newton methods to optimize Browne's (Browne et al. 2002; M. Browne and Cudeck 1989; Browne and Cudeck 1992; Browne and Cudeck 1993) discrepancy function with an appropriate weight matrix that converges to ML, ULS or GLS solutions for the SEM or to Browne's asymptotically distribution free discrepancy function using tetrachoric correlations. Jöreskog (Joreskog 1970) extended this method to allow a priori specification of factors and factor loadings (i.e., the covariance of an unobserved factor and some observed 'indicator') calling this confirmatory factor analysis. Overall fit of the a priori theorized model to the observed data could be measured by likelihood ratio techniques, providing a powerful tool for theory confirmation.

In work that paralleled Jöreskog's, Herman Wold (Wold 1966) described a procedure to compute principal component eigenvalues by an iterative sequence of OLS regressions, where loadings were identical to closed-form algebraic methods. In his approach the eigenvalues can be interpreted as the proportion of variance accounted for by the correlation between the respective 'canonical variates' (i.e., the factor loadings) for weighted sum scores of the two sets of variables. These canonical correlations measured the simultaneous relationship between the two sets of variables, where as many eigenvalues are computed as there are canonical roots (i.e., as many as the minimum number of variables in either of the two sets). Wold showed that his iterative approach produced the same estimates as the closed form algebraic method of Hotelling(Hotelling 1936) , and then through (Lohmoller 1988; Lohmöller 1989) to PLS-PA computer software which generated a sequence of canonical correlations along paths on the network.

The PLS-PA designation has caused no end of confusion in the application of Lohmöller's software, which was casually and somewhat gratuitously called 'partial least squares' as a marketing ploy. Wold was well known for his development of the entirely distinct partial least squares regression (PLSR) NIPALS algorithm. NIPALS provided an alternative to OLS using a design matrix of dependent and independent variables, rather than just the independent variables of OLS. PLSR tends to work well for multicollinear data, but otherwise offers no advantage over OLS. Hauser and Goldberger[(Hauser and Goldberger 1971)(Hauser 1972)(Bielby and Hauser 1977) were able to estimate a model of indicator (observed) variables and latent (unobserved) factors, with correlated indicators and error terms, using GLS; this work provided the basis for Wold's (Hill 1979; Hill 1992) NIPALS algorithm through alternating iterations of simple and multiple OLS regressions. After Herman Wold's death in 1992, PLSR continued to be promoted by his son, the chemist Svante Wold, through his firm Umetrics. Consequently, the most frequent application of PLSR is found in chemometrics and other natural sciences that generate large quantities of multicolinear data.

1.9 Limited and Full Information Methods

The search for estimators for simultaneous equation models can take place in one of two ways: (1) 'limited information' or path methods; and (2) 'full information' or network methods. Limited information methods estimate individual node pairs or paths in a network separately using only the information about the restrictions on the coefficients of that particular equation (ignoring restrictions on the coefficients of other equations). The other equations' coefficients may be used to check for identifiabilty, but are not used for estimation purposes. Full information methods estimate the full network model all equations jointly using the restrictions on the parameters of all the equations as well as the variances and covariances of the residuals. These terms are applied both to observable and to latent variable models. The most commonly used limited information methods are ordinary least squares (OLS), indirect least squares (ILS), two-stage least squares (2SLS), limited information maximum likelihood (LIML) and instrumental variable (IV) approaches. The OLS method does not give consistent estimates in the case of correlated residuals and regressors (which is commonly the situation in SEM analyses), whereas the other methods do provide consistent estimators. Yet OLS tends to be more robust with respect to model specification errors than the other limited information approaches, a problem that is exacerbated by small sample sizes. ILS gives multiple solutions, thus is less favored than other approaches. The 2SLS approach provides one particular set of weightings for the ILS solutions; it also is a particular instrumental variable method. If the equations under consideration are exactly identified, then the ILS, 2SLS, IV and LIML estimates are identical. Partial least squares path analysis (PLS-PA) is also a limited information method. Dhrymes (Dhrymes 1974; Dhrymes et al. 1972; Dhrymes 1972) provided evidence that (similar to Anderson's LIML) PLS-PA estimates asymptotically approached those of 2SLS with exactly identified equations. This in one sense tells us that with well-structured models, all of the limited information methods (OLS excluded) will yield similar results. We will revisit these results when we discuss the behavior of PLS-PA estimators with varying sample sizes in the next chapter.

Chapter 2

Partial Least Squares Path Analysis

Partial Least Squares Path Analysis (PLS-PA) has achieved near cult-like stature within its circle of practitioners, having been touted as the ‘magical bullet’ of statistics (Hair, Ringle, and Sarstedt 2011) (Marcoulides and Saunders 2006) that can find causality where no other method can. In fairness, such extravagant claims are not without their critics. They have in turn incited well-argued rebuttals that PLS-PA is a “voodoo science” (Sosik, Kahai, and Piovoso 2009) and communities today are divided and tribal to a degree unobserved in the past. Many issues arise from PLS-PA not being a proper statistical ‘methodology’ – it has failed to accumulate a body of statistical research on assumptions, the role of data, objectives of inference, statistics or performance metrics. Rather, PLS-PA consists of a half dozen or so software packages that though only lightly documented, seem to be able to conjure path estimates out of datasets that other methodologies reject as inadequate. The exception seems to be Monecke and Leisch (2012a) which has integrated subsequent research into Wold’s mathematics and provided complete documentation on theory and use (though much of this is still controversial). This chapter explores whether PLS-PA software really possesses some ‘secret sauce’ that makes it possible to generates estimates from weak data; or conversely,whether such imputed path structures may indeed be illusory.

Within a limited set of research objectives, PLS-PA is an adequate tool. It’s major failing is that it encourages conclusions based on inadequate, sloppy and flawed data collection. Its major attraction is that it is easy to use, requiring no background in statistics or research design. Sadly these failings have too often been used to justify specious and otherwise indefensible theories. In consequence, PLS-PA has left a trail of junk science and false conclusions that other methods would have avoided.

PLS-PA lacks many of the performance and fit statistics that competing methods offer. When fit statistics do actually exist for the PLS-PA method, they tend to be loosely documented or lack formal statistical development in research papers. The development of PLS path analysis began as a legitimate search for solutions to statistical problems that presented themselves in the 1950s and 1960s. It has been superseded by better methods and most contemporary research disciplines have rejected PLS-PA software as an accepted method for data analysis, despite its practice in a few academic niches. For this reason, the current chapter will try to fill the gap in statistical literature on PLS-PA, while avoiding the risk of legitimizing a controversial and potentially misleading approach to data analysis. The PLS moniker itself is misleading, and has served to confound intellectual boundaries as well as terminology of the PLS culture since its inception in the early 1960s. Hermann Wold developed partial least squares (PLS) regression in the 1950s out of related algorithms he had earlier applied to generating canonical correlations (i.e., correlations between pairs of groups of variables). He also applied his canonical correlation algorithms to latent variable path models; this became known as PLS, even though it did not involve partial least squares regression, and its development was entirely from PLS regression. Nonetheless, it is not uncommon to see PLS articles confound the terminology of path modeling and PLS regression even though the two have nothing to do with each other. This may be used divisively to argue, for example, that path analysis is being used for spectral analysis, when in fact it is regression that is being used (since both are named 'PLS' by the community), or that path analysis is using widely accepted regression methods. This purposeful confusion is compounded by a lack of documentation and supporting research for the software algorithms, a lack of agreement in statistics reported by competing PLS software, and obfuscation by reference to a single 'PLS algorithm'(see Chin and Dibbern (2009), Chin (1998), W.W. Chin (2010a) and Chin and Dibbern (2010)).

2.1 PLS path analysis software: functions and objectives

Wright's path analysis grew in popularity in the 1950s. Researchers in psychometrics, sociology and education were particularly interested in fitting data to models comprised of unobservable quantities such as intelligence, happiness, effort, and so forth. These 'latent' variable path models could not be fit with Pearsonian correlations, rather required more complex underlying modeling assumptions.

Wold (1961) had spent decades developing his algorithms for principal component analysis (where his 'components' could easily be identified with 'latent variables') and with canonical correlation, and proposed that path coefficients be pairwise estimated (concurring with Wright's method) using his canonical correlation algorithm. His particular canonical correlation algorithm was created with the objective of maximizing the correlation between any two latent variables on a path. The overall effect on the model was to significantly overstate the correlation (i.e., the path coefficient) between any two latent variables on the path.

Wold's method was guaranteed to generate a path with significant path coefficients for every data set, since any two variables are likely to be correlated whether or not there really exists any actual causal relationship. This makes it very easy for lazy researchers to 'analyze' sloppy, poorly constructed or misguided datasets, yet find a path structure to support whatever theory is convenient or popular.

2.2 Path Regression

Sewall Wright's path analysis was widely used in genetics and population studies in the first part of the 20th century. During that time, Wright attempted to interest researchers in the fields of psychometrics, survey research, and econometrics in path models, seeing similarities to problems in population studies. Wright's (see Wright (1921), Barclay, Higgins, and Thompson (1995), Wright (1934))original path analysis defined the links between variables as correlations; causal (directional) arrows and specific restrictions on recursive paths were assumed a priori. Wright's widespread popularization of path analysis encouraged statisticians to consider other algorithms for computing path coefficients. During a sabbatical year at the University of Chicago, his path analysis was discussed widely with econometricians who favored regression coefficients. These discussions influenced the subsequent applications of the work of a number of statisticians – in particular Herman Wold and his student Karl Jöreskog.

Econometricians in the 1950s were rapidly developing their field around regression analysis, Tukey Cochran, Mosteller, and Tukey (1954) advocated systematic replacement in path analysis of the dimensionless path coefficients by the corresponding concrete path regressions. Wright (1960) subsequently took to refereeing to these as the 'standardized' and 'concrete' forms of path analysis. The 'concrete' form came to dominate computer intensive path analysis that is used today. Wright (1960) argued convincingly that estimating the concrete form was unnecessarily complex (especially in the days before electronic computers) and that concrete estimators could be recovered from the standardized estimators anyway.

Geneticists Turner and Stevens (see Turner and Stevens (1959)) published the seminal paper presenting the mathematics of path regression (a term coined by Wright). Turner and Stevens' paper ushered in modern path analysis, and their mathematics provided the basis for the computer intensive techniques developed in the 1970s and 1980s.

During the same period, various social science fields – especially psychometrics and education – were investing significant effort in standardizing and quantifying measures of abstract concepts such as intelligence, scholastic achievement, personality traits, and numerous other unobservables that were increasingly important to US national planning and funding in the 1950s. The approach to measuring unobservable (or latent) quantities was essentially to 'triangulate' them by measuring numerous related and measurable quantities. For example, intelligence tests might require an individual to spend several hours answering questions, performing tasks with speed and accuracy, problem solving and so forth in order to assess one unobservable quantity – the intelligence quotient. These problems were more naturally suited for the canonical correlation approaches of Wold and Hotelling than they were for approaches that restricted theorists to observable variables. Wold showed that his iterative approach to computing path correlations produced the same estimates as the closed form algebraic method of Hotelling (Hotelling (1936)). By the late 1970s, Wold's group had implemented his canonical path correlations in Fortran computer software (see Lohmoller (1988), Lohmöller (1989) and Lydtin et al. (1980)) which generated a sequence of canonical correlations along paths on the network. This software came to be called PLS-PA.

The PLS-PA designation has caused no end of confusion in the application of Lohmöller's software, which was casually and somewhat gratuitously called ‘partial least squares’ as a marketing ploy. Wold was well known for his development of the entirely distinct partial least squares regression (PLSR) NIPALS algorithm. NIPALS provided an alternative to OLS using a design matrix of dependent and independent variables, rather than just the independent variables of OLS. PLSR tends to work well for multicolinear data, but otherwise offers no advantage over OLS.

Controversies surround the various interpretations of coefficients. The coefficients for any given model that are generated by a particular software package are likely to diverge significantly from those computed by an alternative software package. This has created problems for interpretation and even defense of construct validity, which have been documented in numerous studies (see K. Bollen (1989), Henseler and Fassott (2010), Henseler, Ringle, and Sinkovics (2009),Lohmoller (1988),McArdle and Epstein (1987),McArdle (1988),Ringle (n.d.),Ringle, Sarstedt, and Straub (2012),Ringle and Schlittgen (2007),Ringle, Wende, and Will (2010),Tenenhaus et al. (2005), Tenenhaus, Amato, and Esposito Vinzi (2004), Bastien, Vinzi, and Tenenhaus (2005)).

2.3 Hermann Wold’ Contributions to Path Anaysis

Hermann Wold brought two important innovations to path analysis: (1) latent variables which he conceived as principal components of the indicators; and (2) a widely used tool to estimate path regression coefficients (versus Wright’s earlier correlation coefficients). Further, a large bit of the innovation in path models evolved around Hermann Wold’s work in the 1950’sand 1960s. Hotelling’s (Hotelling (1936)) seminal work in principal components analysis (PCA) proposed an algorithm to compute the first principal component as a weighted linear composite of the original variables with weights chosen so that the composite accounts for the maximum variation in the original data. Wold’s work in the 1940s and 1950s improved on Hotelling’s computational methods. His work led eventually to regression algorithms for principal components regression (PCR) and partial least squares regression (PLSR) which computed regression coefficients in situations where data was highly multicolinear.

As a byproduct of this work, he started promoting the principal components as representations of abstract (latent) constructs. Latent abstractions proved useful in the evolving fields of psychometrics and sociological surveys, and were widely adopted in the 1950s and 1960s. Social scientists need to model many abstract and unobservable constructs – things like future intentions, happiness, customer satisfaction, and so forth. Though indirectly observable, there were numerous surrogates that could provide insights into such abstract (or latent) constructs – these observable surrogates were called ‘indicators’ of the latent variable.

Wold helped this evolution along by proposing modifications to Wright’s path analysis in the following fashion:

1. Let the research choose indicators for each latent construct in advance of any statistical analysis

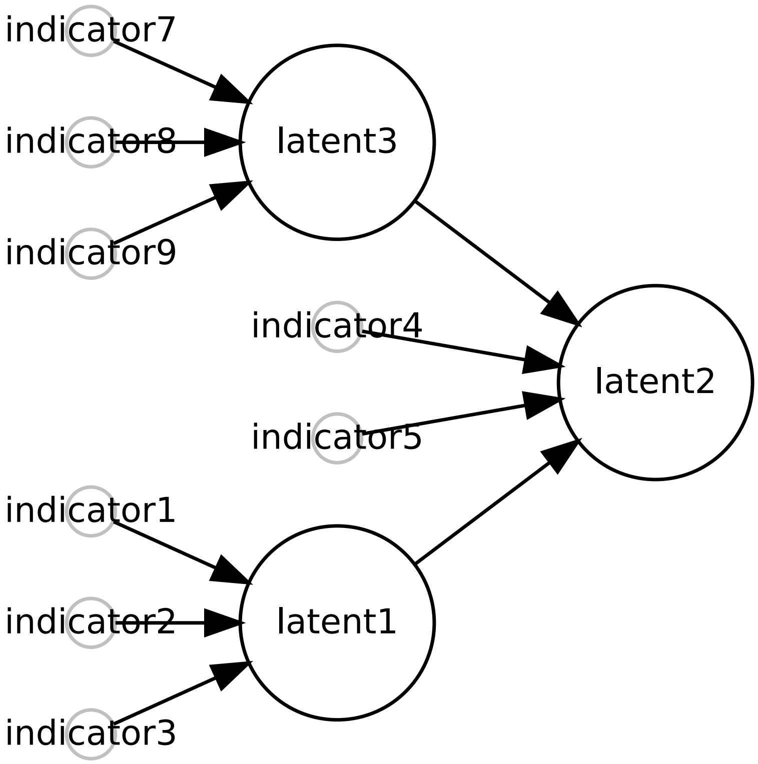

Figure 2.1: PLS-PA Inner and Outer Models

2. Compute the first principal component of each cluster of indicators for a specific latent variable
3. Construct a synthetic latent variable equal to the sum of indicator value multiplied times factor weights from the first principal component for each observation.
4. Compute a regression coefficient between each pair of latent variables in the model using either an OLS or a PLSR regression on the first PCA components of the treatment and response (i.e. tail and head of the link arrow) latent variables. In the OLS case Wold called PCA-OLS setup a principal components regression PCR. Unless the correlations between any two variables are greater than 0.95, both methods produce nearly the same coefficient.
5. Compute each regression coefficient around the model following the link arrows following Wright's three path laws
6. Outer and inner models
    a. The network model of latent variables is called the 'structural model' or sometimes the 'inner' model. The term 'structural equation model' came about from this use, which Wold borrowed from the matrix terminology of systems of equation regression approaches developed at the Cowles Commission.
    b. The clusters of indicators for each latent variable (with links being the factor weights) are sometimes called the 'outer' model

7. Formative and reflective links

    a. If the researcher chooses the indicators before the latent variables, the links are called 'formative' because the factor (link) weights and the latent variable are formed from the first component of the PCA

    b. If the researcher chooses the latent construct before the indicators, the links are called 'reflective' because the factor (link) weights are believed to reflect the abstract (latent) construct. This belief must be validated by reviewing the first component of the PCA, usually through a statistic like Cronbach's alpha.

2.4 Possible Choices for Path Coefficients: covariance, correlation and regression coefficients

2.5 Covariance and Variance

Following Cochran, Mosteller, and Tukey (1954) path modeling adopted a more colorful palette of path metrics, incorporating covariances, variances, and regression coefficients, as well as correlations, which were Sewall Wright's preference for path coefficients (see (Wright, 1960)). Prior to surveying the detailed methods which comprise modern path analysis, it would be beneficial to recap the interpretation of each of these particular measures.

Variance is the second central moment about the mean of a single variable. The square root of the variance is the standard deviation, and provides a linear measure of the variation in that variable.

Covariance is a measure of how much two random variables change together. If the variables tend to show similar behavior, the covariance is a positive number; otherwise if the variables tend to show opposite behavior, the covariance is negative. The sign of the covariance describes the linear relationship between the variables. The magnitude of the covariance is difficult to interpret; thus correlation is typically a better statistic of magnitude of behavioral relationships.

2.6 Correlation

Correlation is the normalized version of the covariance, obtained by dividing covariance by the standard deviations of each of the variables. Correlation ranges from . Several different formulas are used to calculate correlations, but the most familiar measure is the Pearson product-moment correlation coefficient, or Pearson's correlation. Correlations are simple to interpret and to compare to each other because of their normalized range. Correlations between unobserved (latent) variables are called canonical (for continuous data) or polychoric (for ordinal data) correlation. Correlations provide useful summarizations of large datasets

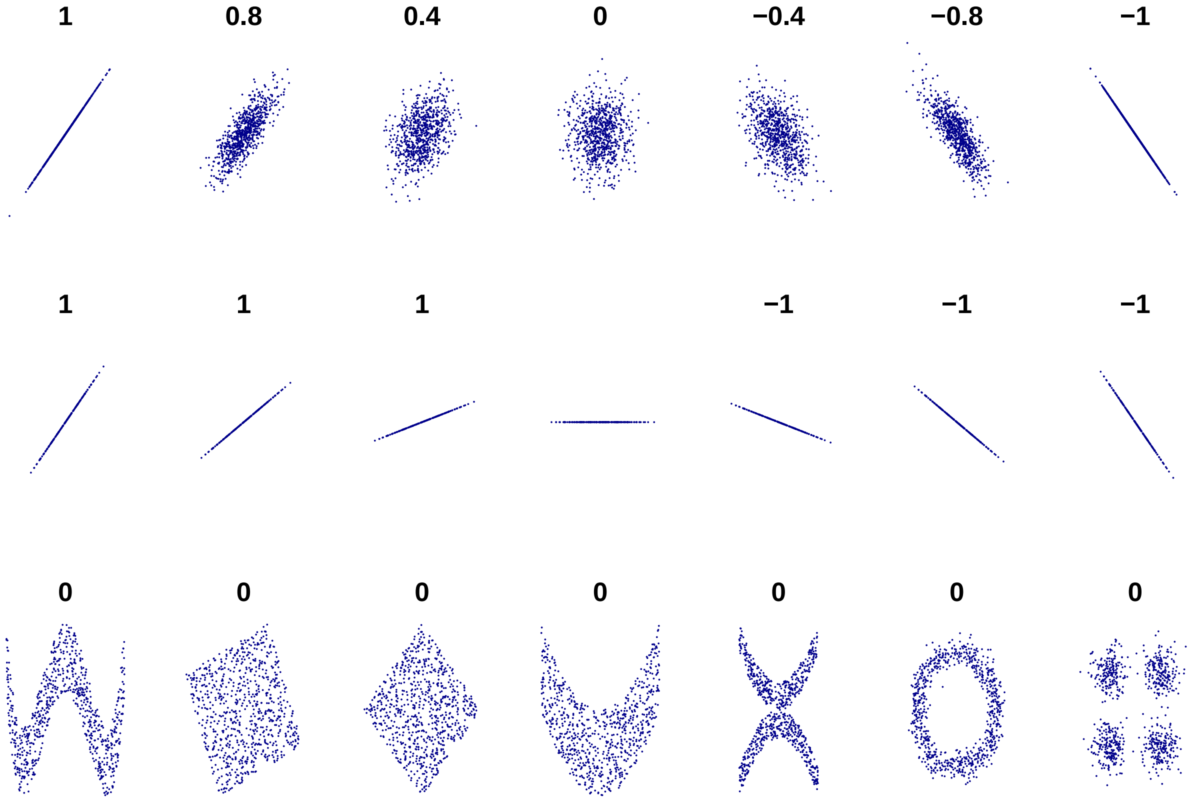

Figure 2.2: The danger of linear assumptions: random variable pairs and their Pearsonian correlations

into single metrics, but the figure illustrates how misleading such extreme summarizations can become.

```
## NULL
```

2.7 Regression coefficients

Regression coefficients provide another measure of the way in which two random variables change together. Many differing approaches to regression are available, each with differing fit objectives, and often involving more than just two variables. Thus there is no universal definition of a regression coefficient. But generally a regression coefficient describes the relationship between a single predictor variable and the response variable – predictor and response imply a causality, with the predictor causing a predictable (based on the regression coefficient) response when all the other predictor variables in the model are held constant.

A regression coefficient (often referenced with the Greek letter) is interpreted as the expected change in the response variable (in some set of measurement units) for a one-unit change in the predictor when all of the other variables are held fixed. Regressions commonly assume linear relationships, but such assumptions may be highly misleading (as are correlations,

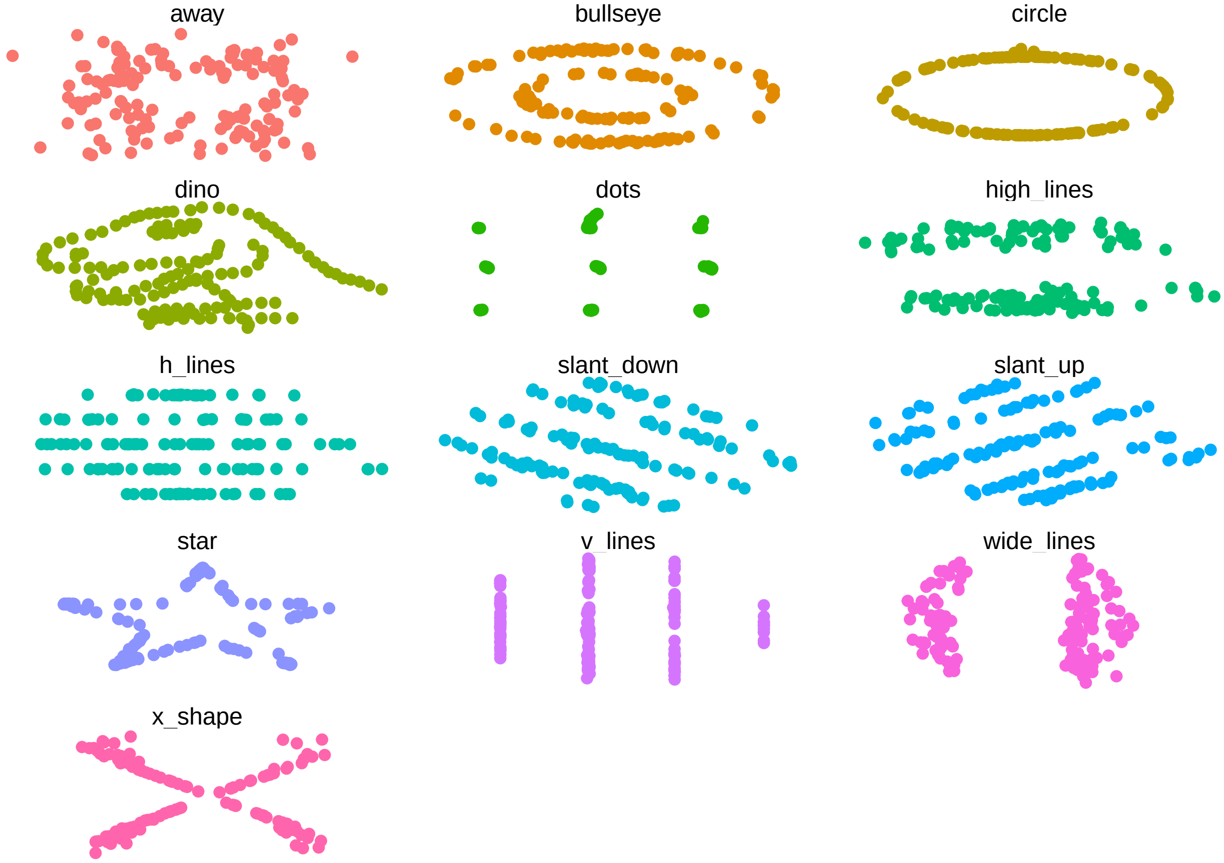

Figure 2.3: (#fig:d_saur)Alberto Cairo's Datasaurus

which are also inherently linear)

dataset	mean_x	mean_y	std_dev_x	std_dev_y	corr_x_y
away	54.266	47.835	16.770	26.940	-0.06413
bullseye	54.269	47.831	16.769	26.936	-0.06859
circle	54.267	47.838	16.760	26.930	-0.06834
dino	54.263	47.832	16.765	26.935	-0.06447
dots	54.260	47.840	16.768	26.930	-0.06034
high_lines	54.269	47.835	16.767	26.940	-0.06850
h_lines	54.261	47.830	16.766	26.940	-0.06171
slant_down	54.268	47.836	16.767	26.936	-0.06898
slant_up	54.266	47.831	16.769	26.939	-0.06861
star	54.267	47.840	16.769	26.930	-0.06296
v_lines	54.270	47.837	16.770	26.938	-0.06945
wide_lines	54.267	47.832	16.770	26.938	-0.06658
x_shape	54.260	47.840	16.770	26.930	-0.06558

Dijkstra (Dijkstra (1983)) p. 76 observed that Herman Wold was generally skeptical of covariance structure methods, because of their assumption of Normal datasets: “Wold questions the general applicability of LISREL because in many situations distributions are unknown or suspected to be far from Normal. Consequently it seemed reasonable to abandon the maximum likelihood approach with its optimality aspirations and to search for a distribution-free, data-analytic approach aiming only at consistence and easy applicability.”

Dijkstra (1983) notes that after developing the PLSR algorithms, ‘Wold and affiliated researchers as Apel, Hui, Lohmöller and S. Wold were, and at present still are, almost exclusively concerned with the development of various algorithms.’ Lohmöller’s algorithms were the most advanced and as we have seen, have formed the basis for modern PLS Path Analysis packages.

Empirical researchers begin (and may too often end) their PLS path modeling data analysis with one of several software packages available. Wold’s work predated modern desktop computers, and the first widely used computer implementation of path regression only appeared in the late 1980s. With the exception of SmartPLS and SPAD-PLS, all of the other PLS path analysis software uses Lohmöller’s software code. Fortunately thorough investigations of software methodologies (see Tenenhaus et al. (2005) , Temme, Kreis, and Hildebrandt (2006) and Monecke and Leisch (2012a)) have been published that provide insight into the internal operations of these packages.

Table 2.1: PLS-PA Software

PLS.Software	Description
LVPLS (Lohmöller 1987)	DOS-based program (Lohmoller, 1988) with two modules for estimating paths: LVPLSC analyzes the covariance matrix, and LVPLSX uses a hybrid PCA-OLS method.. Results are reported in a plain text file. The program offers blindfolding and jackknifing as resampling methods.
PLS-GUI (Li 2005)	Windows-based graphical user interface (GUI) wrapper for Lohmöller’s LVPLS code. It uses the covariance structure analysis LVPLS, and has more in common with approaches like as the first LISREL version in the early 1970s . PLS-GUI produces the same output as LVPLS.
VisualPLS (Fu 2006a)	Another GUI wrapper for Lohmöller’s LVPLS PCA-OLS method. It supports graphics in a pop-up window. Based on the graphical model, the program produces a separate LVPLS input file, which is run by LVPLSX (pls.exe). Various resampling methods are provided.

Table 2.1: PLS-PA Software *(continued)*

PLS.Software	Description
PLS-Graph (Chin 2003)	Another GUI wrapper for Lohmöller's LVPLS PCA-OLS routine (LVPLSX). A limited set of drawing tools are provided for the path diagram, which generates a proprietary input file which cannot be processed by LVPLS. Results are provided in a reformatted LVPLS output. Various resampling methods are provided.
SPAD-PLS (Test&Go 2006)	SPAD-PLS is part of the data analysis software SPAD offered by the French company Test&Go. Models are specified with a menu or graphically in a Java applet. This is the only available software package which actually uses PLS regression,. Transformations of latent variables (squares, cross-products) can be specified. Various resampling methods are provided.
semPLS	The semPLS R-language package is the most professionally documented PLS package currently in existence, with perhaps the most complete and honest exposition of a PLS path analysis algorithm available. It also benefits from the full contingent of R tools, packages and language for pretesting, graphics, and fitting. (Monecke,2012)
SmartPLS (Ringle 2005)	The SmartPLS "drag & drop" interface is the best of the commercial PLS packages, uses OLS and FIMIX algorithms. Various resampling methods are provided. The authors conduct an active and informative user forum at their website.
plspm	R language package dedicated to Partial Least Squares (PLS) methods (CRAN,plsmodeling.com) (Gaston Sanchez, 2012), A corresponding book titled "PLS Path Modeling with R" can be downloaded.
pathmox	R package dedicated to segmentation trees in PLS Path Modeling (CRAN)
qgraph	Network representations of relationships in data (CRAN)

Table 2.1: PLS-PA Software *(continued)*

PLS.Software	Description
psych	Procedures for Psychological, Psychometric, and Personality Research (CRAN)

2.8 Lohmöller's PCA-OLS Path Analysis Method

Of particular importance in the commercial software packages is Lohmöller's PCA-OLS (what Wold had termed principal components regression or PCR) implementation of path regression coefficient estimation. This is used as the default estimation method in all of the commercial PLS Path Analysis software packages, though in fact it is completely different than the PLSR estimation originally envisioned by Herman Wold. Lauro & Vinzi (Vinzi, Lauro, and Amato (2005)) and Chatelin, Vinzi, and Tenenhaus (2002) provide detailed descriptions of Lohmöller's PCA-OLS methodology, though Vinzi, Lauro, and Amato (2005) complain that Lohmöller's LVPLS 1.8 "is only available in a DOS version. It presents important limitations for the number of observations and it is of a very difficult use, completely inadequate for an industrial use." They are probably describing the motivation for the plethora of GUI-wrappers for LVPLS that were developed and sold independently, and now constitute the most commonly used PLS Path Analysis packages.

The Lohmöller's PLS path analysis algorithm can approximated with the following steps:

1. cluster all of the indicator variables into latent variable groupings – either judgmentally (based on the intentions of the survey questions, for example) or using principal components analysis
2. define each latent variables as linear function of indicator variables by assigning factor loadings to each indicator-latent variable link;
3. choose the first component of a PCA on each cluster of factors to define the factor loadings
4. pair-wise regress each latent variables linear combination of factor weighted indicators; this is the path coefficient
5. repeat this procedure following Wright's path diagram constraints until all paths are estimated. Continue to cycle through the model until the estimates converge.

Lohmöller presents these in terms of the 'outer' models – the first principal component of the indicator variables assigned to a latent variable – and 'inner' model – the sequence of OLS path regressions on these latent variables. (Monecke and Leisch 2012b) detail the mathematics of the estimation process, and describes the criterion for stopping when error drops below a predefined critical value. PLS-PA algorithms tend converge very quickly, unless that data carries very little information about the inner model.

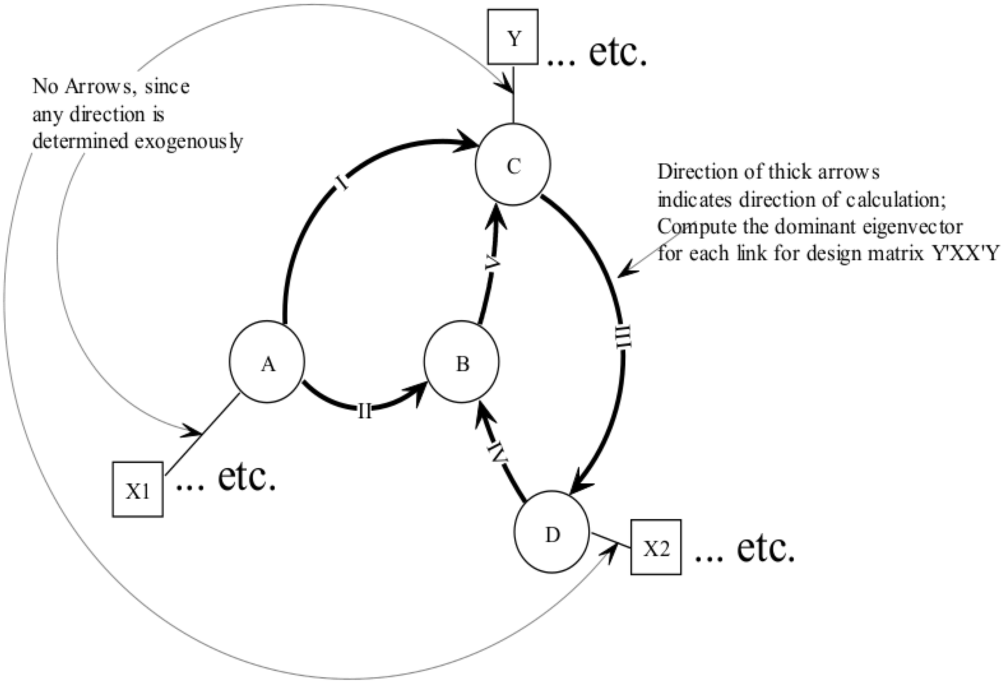
Y
... etc.
No Arrows, since
any direction is
determined exogenously
C
Direction of thick arrows
indicates direction of calculation;
Compute the dominant eigenvector
for each link for design matrix Y'XX'Y
I
V
III
A
B
II
IV
X1
... etc.
D
X2
... etc.

Figure 2.4: Pairwise calculation of path coefficients in Lohmöller's PLS path analysis algorithm

2.9 PLS Path Analysis vs. PLS Regression

Software vendors have created a significant confusion about the methodology they use by including "PLS" in their labeling. Though Herman Wold's original intention was to implement Wright's path regression with partial least squares regression, the implementations created by Jöreskog and Lohmöller in the 1980s did not use partial least squares regression, rather applied two alternate approaches – (1) covariance structure modeling; and (2) a hybrid principal components analysis and ordinary least squares (PCA-OLS) method. Herman Wold himself contributed to the naming confusion, attributing various terms and methods to the acronym PLS.

A brief review of partial least squares regression (PLSR) and principal components regression (PCR) is needed at this point. PLSR and PCA algorithms both estimate the values of factor loadings that help them achieve their statistical objectives. In PCA, that is the unrestricted choice of a set of latent variables (called principal components in PCA) that maximize the explanation of sample variance. PLSR is restricted to maximize the covariance between predictor and response latent variables, fundamentally assuming that the latent variables being studied are structured in a causal model – i.e. the structural equation model of unobserved variables – that has dichotomized latent variables predictors and responses.

In theory, PLSR should have an advantage over PCR. One could imagine a situation where a minor component in independent variables is highly correlated with the dependent variable(s)

; not selecting enough components would then lead to very bad predictions. In PLSR, such a component would be automatically present in the first component (latent variable). In practice, however, there is hardly any difference between the use of PLSR and PCR; in most situations, the methods achieve similar prediction accuracies, although PLSR usually needs fewer latent variables than PCR. Both behave very similar to ridge regression (see Frank and Friedman (1993))

The name partial least squares regression is itself misleading, a fact that was apparent to Herman Wold when he first introduced it in the 1960s. Wold (1966) and Hill (1979) described a procedure to compute principal component eigenvalues by an iterative sequence of OLS regressions, where loadings were identical to closed-form algebraic methods. This was the origin of the term partial least squares, to describe the iterative OLS regressions used to calculate principal components. In his approach the eigenvalues were interpreted as the proportion of variance accounted for by the correlation between the respective 'canonical variates' (i.e., the factor loadings) for weighted sum scores of the two sets of variables. These canonical correlations measured the simultaneous relationship between the two sets of variables, where as many eigenvalues are computed as there are canonical roots (i.e., as many as the minimum number of variables in either of the two sets). Wold showed that his iterative approach produced the same estimates as closed form algebraic method of computing the cross-covariance matrices in canonical correlation analysis.

The name partial least squares regression itself created confusion, and Wold tried in the 1970s to drop regression from the name, without success. As a consequence, at various times, suggestions have arisen to compute regression goodness-of-fit statistics for PLS path analysis, such as R-squared, F-statistics and t-statistics; indeed, PLS computer packages may sometimes even report such fit measures. These numbers exist for individual paths, but are meaningless for the model as a whole.

The authors of the SPAD-PLS software package described in the previous section did take Wold at his word, and does apply PLSR (as an option) to computing path regression coefficients. Thus it is worthwhile at this juncture to elaborate on the methods developed by Wold and justify why these might be desirable in a path model.

In well controlled surveys OLS may be substituted for PLSR. This fact has not been lost on developers of PLS path analysis software. Several studies (see Bastien, Vinzi, and Tenenhaus (2005), Wehrens and Mevik (2007), Temme, Kreis, and Hildebrandt (2006) and Bastien, Vinzi, and Tenenhaus (2005)) reviewed the algorithms used in existing PLS Path Analysis software, and investigated estimation performance. A variety of estimation procedures – including OLS and PLSR – were applied across the packages; Temme, Kreis, and Hildebrandt (2006) notes that "SPAD-PLS is the only program which takes this problem into account by offering an option to use PLSR in the estimation." The general inclination is to apply OLS to the paths, and the factor loadings, path coefficients and R^n are essentially what one would get by regressing the first principal component – a one latent variable principal component regression.

Table 2.2: Strengths and Weaknesses of PLS

Strengths	Weaknesses
PLS Path Analysis is able to model multiple dependents as well as multiple independent variables	PLS Path Analysis software generally uses OLS regression methods (the ‘PLS’ in the name is a misnomer initiated by Wold) thus provides no better estimates than traditional sequential limited information path analysis
PLS Path Analysis software produces estimates from small datasets (though heavy use of sampling makes these estimates difficult to interpret or verify)	The small-sample properties of PLS Path Analysis software are not inherent in the regression algorithm, rather result from intensive, and often poorly modeled and justified resampling of sample data.
PLS predictions are able to handle multicolinearity among the independents;	PLS is less than satisfactory as an explanatory technique because it is low in power to filter out variables of minor causal importance (Tobias, 1997). PLS estimator distributional properties are not known, thus the researcher cannot assess ’fit’, and indeed the term probably is not meaningful in the PLS context
Because PLS Path Analysis estimates one path at a time, models do not need to be reduced and identified, as in systems of equation regression models	PLS Path Analysis is often used to process Likert Scale data – which may be considered either ordinal or polytomous Rasch data. Heavy use of Gaussian resampling is used to force estimator convergence in the software algorithm, which makes assessment of the validity of coefficients difficult.
Heavy reliance on resampling allows PLS Path Analysis to compute estimates in the face of data noise and missing data	Theory-driven introduction of prior information into the resampling distribution models and testing is questionable, because the ‘data are not able to speak for themselves’
in contrast to LISREL SEM. A model is said to be identified if all unknown parameters are identified.	PLS estimator distributional properties are not known, thus the researcher cannot assess ’fit’, and indeed the term probably is not meaningful in the PLS context.

2.10 Resampling

All of the PLS Path Analysis software packages have touted their ability to calculate coefficients from datasets that would be too small for covariance structure modeling methods, or other commonly used statistical methods. Unfortunately, this claim is somewhat misleading, as the software accomplishes this through computer intensive resampling – estimating the precision of sample statistics (medians, variances, percentiles) by replicating available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping). Such techniques are controversial, and it is not my intention to wade into those debates.

Resampling should generally be avoided for another reason. Assume that statistical precision is another way of specifying the Fisher Information in the sample. Then a dataset that is resampled from a given set of observations has exactly the same Fisher Information as the original dataset of observations. The only way that the reported statistics will improve is if the researcher (erroneously) assumes that the additional observations are not resampled, but are new, independent observations. The honest way to collect more information is to increase the sample size.

2.11 Measures

PLS path analysis does not 'fit' the entire model to a data set in the sense of solving an optimization subject to a loss function. A PLS path regression is simply a disjoint sequence of canonical correlations of user defined clusters of indicator variables. As a consequence, fit measures of the whole structural model are not meaningful – the only fit statistics or performance measures possible are ones that apply to individual links. Even the interpretation of standard OLS regression fit statistics on individual links , such as R-squared, F-statistics or t-statistics is confounded by the synthetic nature of the latent variables.

Despite this, numerous *ad hoc* fit measures have been proposed for PLS path analysis, and many may be misleading (see Chin and Dibbern (2009), W.W. Chin (2010a), Chin (1998), W.W. Chin (2010b), Henseler and Fassott (2010), Henseler, Ringle, and Sinkovics (2009), Ringle and Schlittgen (2007), Ringle (n.d.), Ringle, Wende, and Will (2010) for examples of various ad hoc metrics with discussion). Some of these 'fit' measures have arisen from a confusion about what PLS actually does. Wold (see Hill (1979), Wold (1966), Hill (1979))) finessed the role of PLS with explanations of 'plausible causality' – a hybrid of model specification and exploratory data analysis – which added to rather than resolving confusion over his innovation.

2.12 Limited Information

PLS path analysis is often called a 'limited information' method without bothering to further define what exactly this means. But this is precisely where the lack of fit statistics generate the greatest impact on PLS path analysis.

Limited information in path analysis implies that OLS estimators on individual pairwise paths will, in most practical circumstances, replicate the results of PLS path analysis software. But in specific ways, separate regressions could improve on PLS path model software programs. With individual regressions, R-squared, F-statistics and residual analysis through graphing and other tests can reveal a great deal of information about the underlying data. PLS path analysis software too often obscures such information in the data by (1) resampling, which imposes prior beliefs about the population on the data; and (2) overreaching by claiming to assess the entire model at once. Research is generally better served by a full information method such as covariance approaches (e.g., LISREL, AMOS) or a system of equations approach.

2.13 Sample Size in PLS-PA

A stubborn mythology surrounds notions of data adequacy and sample size in PLS-PA that originated with an almost offhand comment in Nunnally (1967) who suggested (without providing supporting evidence) that in estimation 'a good rule is to have at least ten times as many subjects as variables.' Justifications for this rule of 10 appear in many frequently cited publications (see Barclay, Higgins, and Thompson (1995),Chin and Dibbern (2009),Chin (1998),Chin and Dibbern (2009),W.W. Chin (2010a),Chin and Dibbern (2010),W.W. Chin (2010b),Chin and Newsted (1999) and Kahai and Cooper (2003)). But Goodhue, Lewis, and Thompson (n.d.),Boomsma (1982),Boomsma (1985),Goodhue, Lewis, and Thompson (n.d.)„Boomsma (1987), Ding, Velicer, and Harlow (1995) and others have studied the rule of 10 using Monte Carlo simulation, showing that the rule of 10 cannot be supported. Indeed, sample size studies have generally failed to support the claim that PLS-PA demands significantly smaller sample sizes for testing than other SEM methods.

Adaptive bias becomes a significant problem as models grow more complex. Rather than seeking absolute truth or even rational choice, our brains adapt to make decisions on incomplete and uncertain information input, with an objective of maximum speed at minimum energy expenditure (see Gilbert (1998), Henrich and McElreath (2003) , Neuhoff (2001). This manifests itself in anchoring and adjustment on preconceived models (Kahneman & Tversky, 1979 and stronger and stronger biases towards false positives as models grow more complex. The latter problems might be considered one of intellectual laziness, as the brain does not want to expend effort on thinking about an exponentially increasing number of alternative models.

Statisticians use the term 'power' of a test to describe the ability of a method to test a particular model and dataset to minimize false positives. Complex network models drive the power of tests towards zero very quickly. Where multiple hypotheses are under consideration, the powers associated with the different hypotheses will differ as well. Other things being equal, more complex models will inflate both the type I and type II errors significantly (Bland and Altman (1995)).

The demand to justify PLS-PA models has generated some strange research curiosities. Citation analysis reveals the centrality of Jum Nunnally in the fields of marketing, organizational behavior and management information systems. Nunnally was a psychometrician at Vanderbilt University and the University of Illinois who lived before the establishment of any of these disciplines, either in practice or academe. He certainly never studied computers or their managerial implications. What he did was coauthor a book on psychometrics in the 1960s (Nunnally, Bernstein, and others 1967). Buried deep in the book was an off-hand remark, unsupported and probably inconsequential, that for estimation 'a good rule is to have at least ten times as many subjects as variables.' Nunnally statistical dogma' as the 'rule of 10' which supposedly allows statistically insignificant samples of questionnaires to generate statistically significant results about unrelated, unobserved, abstract concepts. Nunnally is cited so often as justification for inadequate sample sizes that he has become a top 10 researcher in these fields (Hsu, Westland, and Chiang 2015).

Nearly every study investigating Nunnally's 'rule of 10' for sample size has found it to yield sample sizes that are many orders of magnitude too small. Studies relying on these small samples are unreliable with significantly inflated type I and II error rates. Various studies reported in K. Bollen (1989) and Hu and Bentler (1999), Bentler (1990) and Bentler and Mooijaart (1989) rejected the rule of 10 as fiction, and suggested a possible 5:1 ratio of sample size to number of free parameters. Monte Carlo studies conducted in the 1980s and 1990s showed that SEM estimator performance are not linearly or quadratically correlated with the number of parameters (see M. Browne and Cudeck (1989), Browne and Cudeck (1992), Browne and Cudeck (1993), Gerbing and Anderson (1988) and Geweke and Singleton (1981)). Going further, Velicer et al. (1998) reviewed a variety of such recommendations in the literature, concluding that there was no support for rules positing a minimum sample size as a function of indicators. They showed that for a given sample size, a convergence to proper solutions and goodness of fit were favorably influenced by: (1) a greater number of indicators per latent variable; and (2) a greater saturation (higher factor loadings). Several studies, H. Marsh and Bailey (1991), Marsh, Byrne, and Craven (1992),Marsh and Yeung (1998), Marsh, Wen, and Hau (2004) and Marsh and Yeung (1997) concluded that the ratio of indicators to latent variables rather than just the number of indicators is a substantially better basis on which to calculate sample size, reiterating conclusions reached in Boomsma (1982), Boomsma (1985) and Boomsma (1987). We will revisit this problem later, and provide criteria for control of error inflation and adaptive bias in sample selection for structural equation models.

The availability of PLS-PA software packages allows a relatively straightforward Monte Carlo exploration of statistical power and sample size in PLS-PA. PLS-PA is implemented in around a half dozen or so software packages. Fortunately Monecke and Leisch (2012a) have provided an accessible implementation on the R-language platform (semPLS) that can be used to explore its otherwise arcane characteristics. As this section will show, there are curious idiosyncrasies of PLS-PA that set it apart from widely used statistical approaches. We can show that: 1. contrary to the received mythology, PLS-PA path estimates are biased and highly dispersed with small samples; 2. sample sizes must grow very large to control this bias and dispersion with $bias \propto \frac{1}{samplesize}$ and $dispersion \propto \frac{1}{log(samplesize)}$; and 3. the power and significance of PLS-PA hypothesis test is roughly the same as for 2SLS models, concurring with Dhrymes (1972), Dhrymes (1974), Dhrymes and Erlat (1972), Dhrymes et al. (1972) and for small samples is low at most effect levels, yielding an excessive number of false positives.

Various studies in Hair et al. (2012),Henseler, Ringle, and Sinkovics (2009),Ringle, Sarstedt, and Straub (2012),Ringle (n.d.),Ringle and Schlittgen (2007) and Ringle, Wende, and Will (2010) have reviewed the use of (J. B. Lohmöller, 1981; Lohmoller, 1988; J.-B. Lohmöller, 1989) algorithmic methods and estimators in Lohmoller (1988),Lydtin et al. (1980),Lohmöller (1989). Monecke and Leisch (2012a) noted that all of this "PLS" path model software uses the same Lohmoller (1988) algorithm on an ad hoc iterative estimation technique. Ringle, Sarstedt, and Straub (2012) surveyed a subset of studies using PLS-PA for testing, noting that three-quarters of studies justify the application of PLS path analysis software in a single paragraph at the beginning of the data analysis citing PLS' use of either small sample sizes (36.92%) or non-normal data (33.85%) – more than two-thirds of studies violate the standard assumptions made in estimation with PLS-PA. It's supporters claim that PLS-PA allows latitude in sample size and distributions that no other statistical approach is able to offer.

Monecke and Leisch (2012a) describe their implementation of Lohmöller's algorithm, as follows. Assume a given path model; for example $A \to B \to C$. Assume that (A, B, C) are latent (i.e., unobservable) variables that are each comprised of a pair of observable 'indicator' variables: $A = \sum_i \alpha_n A_n$; $B = \sum_i \beta_n B_n$; and $A = \sum_i \gamma_n C_n$ for $i = 1, 2$. The PLS software maximizes a scaled canonical correlation between pairs of variables by iteratively stepping through the model and adjusting factor weights $(\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2,)$. First, on path $A \to B$ weights are set to initial values, say $(\alpha_1, \alpha_2) = (.5, .5)$ and weights (β_1, β_2) are manipulated to maximize the path coefficient (i.e., a scaled canonical correlation) on $A \to B$. The process is repeated for path $B \to C$ keeping the computed weights (β_1, β_2) and choosing (γ_1, γ_2) to maximize the path coefficient on $A \to B$. This is repeated in a cycle $C \to C \to B \to C \to A \to B \to ...$ until the change in path coefficient values is less than some preset value. The fitting of data to the path model only occurs on individual pairs of latent variables – it is piecewise. Researchers call this 'limited information' fitting of the data; meaning that all of the data outside of a particular path is ignored in maximizing a path coefficient. 'Limited information' approaches to path modeling generate highly inflated path coefficients, and many researchers like that as it lowers their workload in proving a theory. When model paths are preselected (for example $A \to B$ and $B \to C$. are preselected out of three possible paths $A \to B$, $B \to C$ and $C \to A$) these 'limited information' approaches have significantly higher probability of statistically confirming the preselected path $A \to B \to B$ than would a method that considered all of the data simultaneously. This is likely why its supporters assume that PLS-PA works well with small sample size and non-normal data that would not yield statistically significant results using standard approaches.

To analyze the behavior of PLS-PA algorithms, an inner (latent variable) model was tested with the R-code semPLS package on data that follows a zero centered Normal distribution, but is expressed as indicator variables that are Likert scaled 1:5 integer values. This is a standard path setup that is used in many social science research papers. Each of the latent variables was measured with three indicator variables that are Likert scaled 1:5 integer values. Every indicator is statistically independent of any others, and the population correlation structure is an identity matrix (complete independence of observed variables). Random data for this study was generated using the Mersenne-Twister algorithm (Matsumoto and

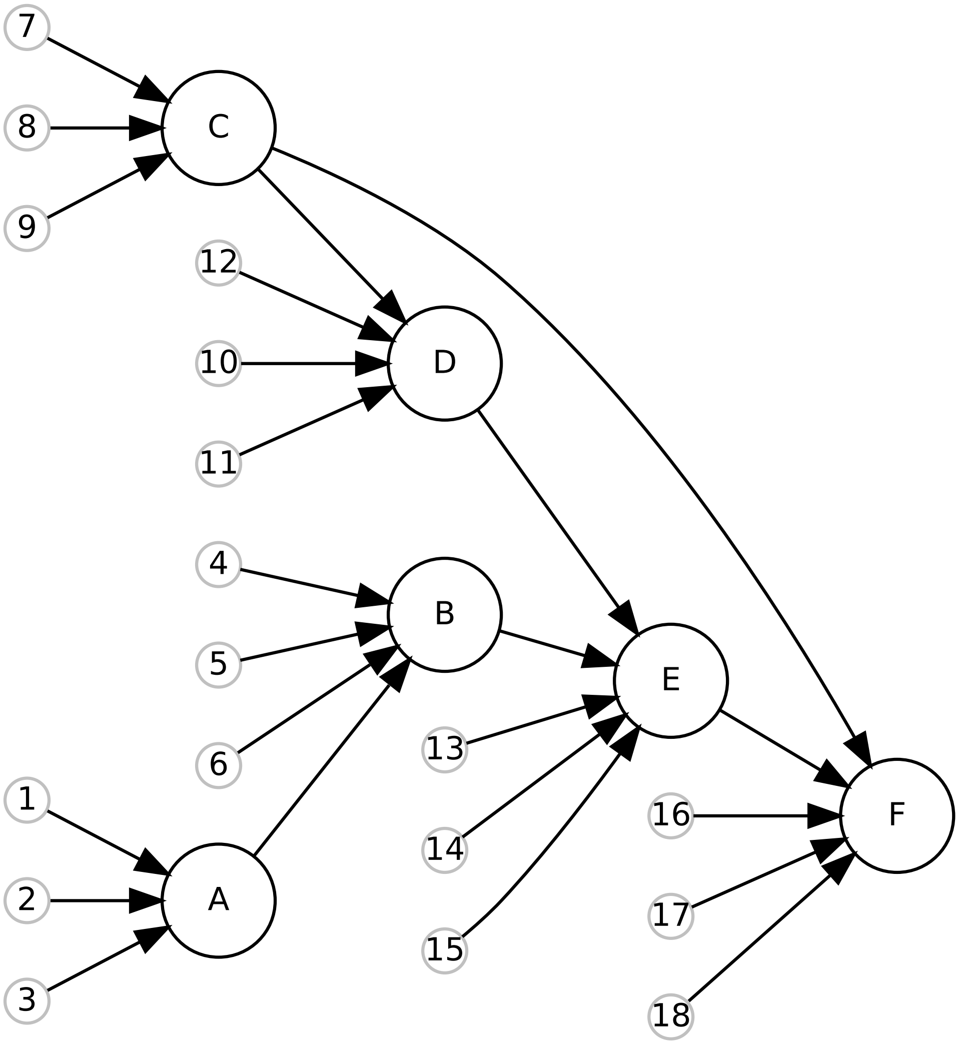

Figure 2.5: The path model analyzed with the semPLS R package

Nishimura (1998)).

The figure 2.6 plots the path estimator standard deviation over 200 trials with sample sizes in the range $[10^2; 10^5]$. Actual correlation was zero (since the sample was random) yet PLS estimators (i.e., canonical correlations) are highly unstable not just for small samples, but even for larger sample sizes. Note that these estimates are for exactly the same input data; samples were simply taken for larger and larger chunks of that data, always starting from the first sample item. Again, it is important to consider that if the PLS-PA algorithm were accurately estimating path coefficients from this population, these values would be zero (since all the observations are independent), and estimators would converge to zero as larger and larger samples were taken. This is clearly not the case.

```r
library("semPLS")
library("ggplot2")
```

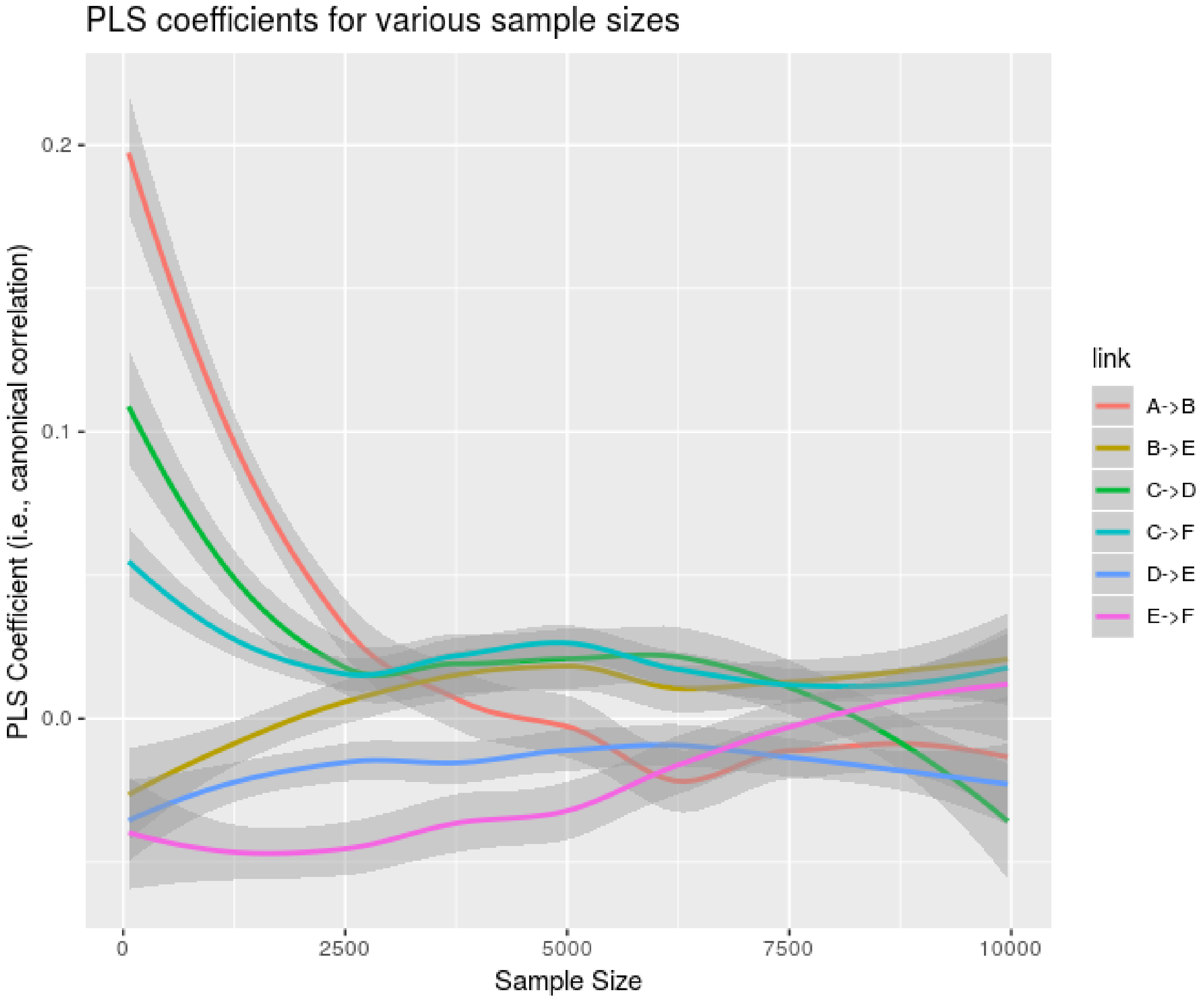

Figure 2.6: Link estimation in PLS for successively larger samples from the same population, showing 95% confidence interval

```
## Set up the structural model

sm <- matrix(c("A" ,"B", "C", "D" ,"B" ,"E", "D", "E" ,"E" ,"F", "C" ,"F"), ncol=2, byrc
colnames(sm) <- c("source","target")

## Set up the measurement model

mm <- matrix(c("A", "A1","A", "A2","A", "A3", "B", "B1", "B", "B2", "B", "B3","C", "C1",
colnames(mm) <- c("source","target")

  ## Populate plsdata with random Likert entries
pcoef <- t(matrix(rep(1,c(6*6)))) ## First row matrix of PLS coefficients
plsdata_full <- floor(1+runif(c(10000*3*6),0,5))
plsdata_full <- data.frame(matrix(plsdata_full,ncol=c(3*6)))
colnames(plsdata_full) <- c( "A1", "A2", "A3", "B1", "B2", "B3", "C1",  "C2", "C3","D1",

## Estimate path coefficients for successively larger samples
for(size in seq(150,10000,50)){
   plsdata <- plsdata_full[1:size,]
   plsmod <- plsm(data = plsdata, strucmod = sm, measuremod = mm)
   plsest <- sempls(model = plsmod, data = plsdata, wscheme = "centroid", verbose=FALSE,

## record the PLS link coefficients (i.e., canonical correlations)
    pcoef <- rbind(pcoef,matrix(plsest$path_coefficients,nrow=1))
}

pc<- data.frame(seq(60,10000,50),pcoef[2:200,13],pcoef[2:200,20],pcoef[2:200,27],pcoef[2
colnames(pc) <- c("x","AtoB","CtoD","BtoE","DtoE","CtoF","EtoF")

library(reshape2)
pcm <- melt(pc, id.vars = 1)
colnames(pcm)=c("sz","link","val")
pcm$link <- gsub("to","->",pcm$link)
pl <- ggplot(pcm)+geom_smooth(aes(sz,val,color=link))+ labs(title="PLS coefficients for
pl
```

This brief example provides us with clear guides on how to interpret the assertions that conflate PLS-PA's ability to generate coefficients without abnormally terminating as equivalent to estimating with small samples. The various software implementations of PLS-PA do generate purported 'goodness-of-fit' metrics such as and so forth, but any fit statistics can only make sense in the limited context of individual path estimates. In the context of a complete path model, these statistics are meaningless. Indeed, Noonan and Wold (1982) alluded to the incompleteness of his approach, referring to it as a 'limited information' approach where path coefficients suggest only 'plausible causalities'. We can glean the following insights from this brief Monte Carlo study:

(1) PLS path estimates are biased and highly dispersed when computed from small samples. Roughly the PLS-PA estimator $bias \propto \frac{1}{sample-size}$ and $dispersion \propto \frac{1}{log(sample-size)}$ but as our example shows, this is neither a smooth or a stable relationship. T

(2) Path coefficient bias and dispersion are significant for the commonly used 5-point Likert scale data. Sample sizes must grow more rapidly than they will in regression for estimating the same data and path models, because the latter have methods for incorporating distributional information (e.g., via general linear models) and analysis of residuals. Exceptional performance when the researcher has collected only small sample sizes or non-normal data is a widely cited justifications for application of PLS path analysis. Yet we see that in fact the PLS-PA algorithm actually requires significantly larger sample sizes to extract information from the population. In effect, the algorithm is throwing away useful data from the sample, but researchers fail to see this, because the software is usually able to compute some number for the path coefficient. Prior literature has conflated PLS software's ability to generate coefficients without abnormally terminating as equivalent to estimating with small samples and non-normal distributions. But this is in fact a flaw in the PLS path analysis software that allows it to generate incorrect results without generating a corresponding warning to the researcher.

(3) The power of PLS hypothesis tests is low across effect levels, leading PLS software to generate a disproportionate number of false positives. In essence, PLS-PA is an ideal p-hacking tool because, even ignoring the cherry picking of data cited by Ioannidis (2005), excessive generation of false positives supports the widespread publication of erroneous conclusions.

Responsible design of software should stop calculation when the information in the data is insufficient to generate meaningful results, thus limiting the potential for publication of false conclusions. Unfortunately, much of the methodological literature associated with PLS software has conflated its ability to generate coefficients without abnormally terminating as equivalent to extracting information and estimation with small non-normal samples. Instead, the current analysis shows that PLS software is generating an inappropriate number of false positives when applied to the commonly invoked models and data types.

2.14 PLS-PA: The Bottom Line

Effort has been expended on making PLS-PA algorithms accessible and usable with survey datasets. The culmination of this effort has arguably been Monecke and Leisch's semPLS package and documentation (Monecke and Leisch 2012b). But at its core, PLS-PA is still factor analysis with some domain specific assumptions (i.e., *a priori* selection of latent constructs) about the underlying problem structure. Indeed, PLS-PA grew out of Wold's research in canonical correlation analysis that optimally describes the cross-covariance between multiple datasets (i.e., the indicators of each latent variable). This allows more flexibility in problem definition than principal component analysis (PCA) which defines a new orthogonal coordinate system that optimally describes variance in a single dataset (i.e., all the indicators combined). But PLS-PA still suffers from the fundamental limitations of PCA when applied to research questions. PCA is a tool for exploratory data analysis, for dimensionality reduction and for model specification search (what is sometimes called predictive modeling).

In its currently computer implementations, PLS-PA measures linear relationships between hypothesized variables given an arbitrary dataset. PLS-PA suffers from deficiencies make it unsuitable for testing research questions:

1. Path coefficients in PLS-PA are sub-optimal canonical correlation measures. Since canonical correlation analysis defines coordinate systems that optimally describe the cross-covariance between datasets, and since the inner PLS-PA model imposes arbitrary constraints on the dataset membership, and these arbitrary constraints will yield path coefficients that cannot explain as much of the data variance as would a true canonical correlation metric.

2. PLS-PA lacks any distance measures beyond a simple linear distance metric. This is important for human subjects studies, because many human perceptual relationships are logarithmic or otherwise non-linear. Non-linear relationships between latent constructs would be missed in a PLS-pa analysis.

3. PLS-PA lacks fit statistics to measure how well the dataset conforms to the researchers hypotheses which are derived from the research question

4. PLS-PA cannot determine causal direction; correlations are inherently non-direction distance measures.

5. Because PLS-PA lacks fit statistics, it is impossible to determine the needed sample size to yield enough information to test a hypothesis. Hypothesis testing is meaningless in the context of PLS-PA.

PLS-PA's distribution-free character is often cited as a justification for using the method. Indeed, a non-parametric, distribution-free method is needed for Likert-scaled data which is both integer-valued and truncated at zero and a small positive integer, something that would be modeled with a multinomial distribution. The problem is that PLS-PA's non-parametric properties derive from the fact that it produces linear distance measures between datasets; it was not intended to validate research models by fitting data to distributions. PLS-PA throws away any distributional information that might be used to construct causal models, instead asking the research to construct the *a priori*.

According to an old superstition, bullets can be charmed to make sure that they would hit a particular person; they are called 'magic bullets' and only a Bible can deflect such projectiles. Sadly, little evidence exists today that either magical projectiles or magical statistical methods are anything more than illusory. If this sounds too pessimistic, it is. PLS-PA can yield estimates comparable to AMOS, LISREL and other full-information SEM methods. But it cannot yield reliable estimates with small or highly multicolinear dataset. Indeed it may be less suitable for weak data because the method provides no feedback on the quality of estimate; it just computes something if it is fed numbers. PLS-PA an ideal tool for unscrupulous or lazy researchers interested in supporting bogus theories with random data. Much of the negative press surrounding PLS-PA arises from this type of abuse.

Although PLS-PA has significant limitations, is simple to setup and use and invariably generates output even with the weakest of data. This makes it especially useful for exploratory analysis and model specification search. Problems arise when it is applied where it should not be: in model evaluation, hypothesis testing and analysis of causal direction.

Chapter 3

Full-information Covariance SEM

Sewall Wright's path coefficients were conceived as dimensionless binary indicators of whether a genetic trait was passed to an offspring, or not; they were 'bits' of information about whether or not a gene was present. Correlations in Wright's context were almost overkill, though their magnitude might have been considered to suggest varying degrees of confidence in heritability of a trait. As path analysis began to find application in analyzing relationships that were multivalued or continuous, limitations in the ability to resolve effects began to reveal themselves. It was in this context that (Cochran, Mosteller, and Tukey 1954) advocated systematic replacement in path analysis of the dimensionless path coefficients by the corresponding concrete path regressions. In the early days of data processing, both Herman Wold and his student Karl Jöreskog developed software, building on Turner and Stevens' mathematics, that took advantage of computer intensive techniques becoming available to universities in the 1970s and 1980s. Jöreskog substantially advanced the utility of SEM by creating the LISREL full-information method for SEM analysis, which allowed fit statistics and with proper argumentation, the testing of causal models and hypotheses on those models.

3.1 LISREL

Wold's student, Karl Jöreskog, extended Wold and Lohmöller's (Lohmöller 1989)(Lohmoller 1988)methods in software implementations of covariance structure path analysis methods. Jöreskog's LISREL (an acronym for LInear Structural RELations) software was the early trend setter in computer intensive path model, having appeared in a mainframe form in the late 1970s. Later in the 1980s (prior to the Windows 3.1 graphical interface) Lohmöller released desktop computing software that implemented both Jöreskog's as well as his own interpretation of Wold's ideas for path modeling in LVPLS. LISREL followed shortly with a desktop version, and these two packages set the standards for future path modeling software. Lohmöller's proponents have projected an ongoing animosity towards the LISREL and AMOS approaches. PLS path analysis is to this day considered to be a competitor of Jöreskog's LISREL approach. (Vinzi, Lauro, and Amato 2005) complained: "The goal of LISREL (or hard modeling) is actually to provide a statement of causality by seeking to find structurally or functionally invariant parameters, i.e. invariant features of the mechanism generating observable variables) that define how the world of interest to the model at hand works. These parameters are supposed to relate to causes describing the necessary relationships between variables within a closed system. Unfortunately, most often real data do not meet the requirements for this ideal." Whether or not you accept Lauro and Vinzi's complaints about LISREL (and covariance structure modeling approaches in general) many researchers tend to be frustrated with the LISREL software for two main reasons:

1. It requires larger sample sizes, and simply does not computer path coefficients at all if the sample size is too small (PLS path analysis software addresses this through the dubious methods of resampling)

2. It requires that data be Normally distributed, which is a problem with survey data, which typically involves a distinctly non-Normal 5 or 7 point Likert scale censored at zero.

Debates aside, Jöreskog initially focused on the corresponding concept for undirected graphs – a forest, or an undirected graph without cycles. Choosing an orientation for a forest produces a special kind of directed acyclic graph called a random tree, for which stochastic process modeling was reasonably well understood. Jöreskog based his confirmatory factor analysis (CFA) methods on random forests, ultimately extending these methods to the full linear structural relations (LISREL) software package introduced in the 1970s. LISREL was originally syntax-based and mainframe based; it was promoted in marketing research by Claes Fornell at the University of Michigan. Jöreskog's innovation was to conceive of path coefficients as covariances, thus one more solution concept to Wold's path models with latent variables comprised of the first components of indicators (i.e., observations). Lohmöller's path analysis software included two methods for computing path coefficients between the latent variables (i.e., first principal components) – ordinary least squares regression; and covariance analysis. Subsequent covariance method software has proven inconsistent in its methods for calculating path coefficients. It has even equivocated on whether path coefficients should be reported as covariances, as would be the product of covariance analysis; or whether these should be presented as correlations, consistent with Wright's original approach.

AMOS, a popular software package, now owned by IBM-SPSS, was developed by (McArdle and Epstein 1987; McArdle 1988). reformulates Jöreskog's mathematics in the more compact RAM format, and reports correlations, thus blurring the bounds between path analysis methodologies. Ultimately, though, as Wright (1960) observed, any particular format for reporting path coefficients can be recovered from the others, and the only important difference is in ease and utility of interpretation. In all cases, though, covariance methods are a full information estimation method (all of the latent variables are simultaneously used in the path coefficient calculations) as opposed to Wright's or Wold's path analysis, which were limited estimation methods that computed the coefficient between each pair of latent variables individually and sequentially. The basic method is to use the indicator values to estimate the covariance matrix. The diagonal of the covariance matrix contains the latent variable's variance, the off diagonal elements, the covariance between two latent variables. Some covariance software report correlations (a la classic Wright path coefficient) which they

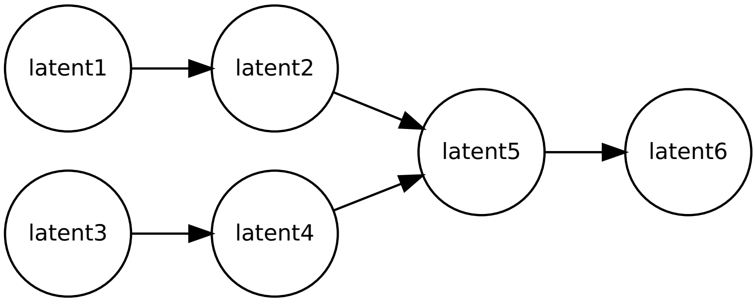

Figure 3.1: Six latent constructs with five causal links

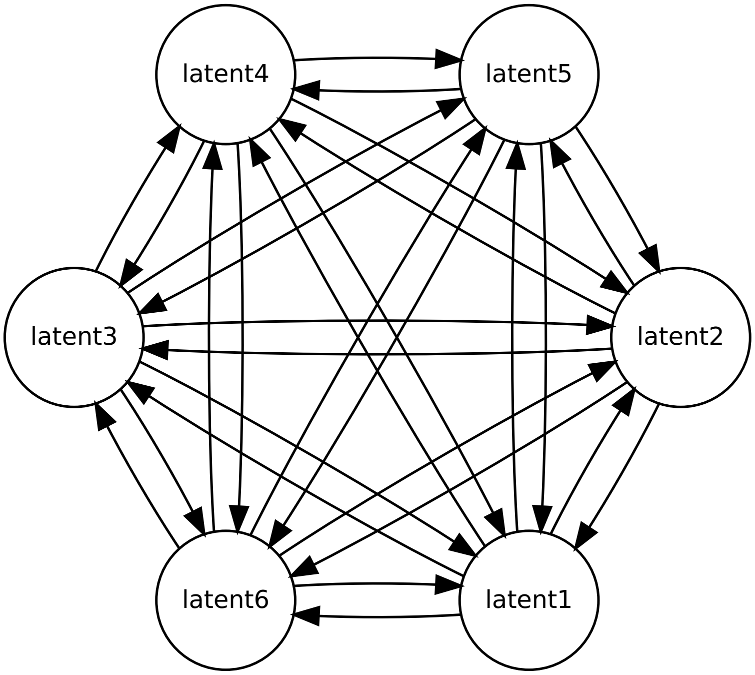

Figure 3.2: Each link can either exist (be significant) or not, yielding 29 competing models

compute by dividing the path correlation by the square roots of the two diagonal terms (one for the row number, and one for the column number).

LISREL, like other SEM methods, has difficulty differentiating between competing models that incorporate the same latent variables. A single causal model, for example one whose links might reflect a specific theory like the Technology Acceptance Model structure shown in figure 3.2 with n latent variables is competing with $n^2 - n - 1$ alternative causal models based on the same set of latent variables.

For discriminant analysis, you would assume that each causal relationship (arrow) points forward, backward or does not exist. A TAM structured model such as shown in figure 3.2 implies that the only 5 entries in the covariance matrix are non-zero; The other 10 entries in the covariance matrix are restricted to be zero valued.

3.2 Short history of LISREL

Jöreskog (Joreskog 1970) developed a rapidly converging iterative method for exploratory ML factor analysis (i.e., the factors are not defined in advance, but are discovered by exploring the solution space for the factors that explain the most variance) based on the Davidon-Fletcher-Powell math programming procedure commonly used in the solution of unconstrained nonlinear programs. As computing power evolved, other algorithms became feasible for searching the SEM solution space, and current software tends to use Gauss-Newton methods to optimize Browne's discrepancy function (see (Browne et al. 2002; M. Browne and Cudeck 1989; Browne and Cudeck 1992; Browne and Cudeck 1993)) with an appropriate weight matrix that converges to ML, ULS or GLS solutions for the SEM or to Browne's asymptotically distribution free discrepancy function using polychoric (i.e. latent variables) correlations. Jöreskog (Joreskog 1970) extended this method to allow a priori specification of factors and factor loadings (i.e., the covariance of an unobserved factor and some observed 'indicator') calling this confirmatory factor analysis. Overall fit of the a priori theorized model to the observed data could be measured by likelihood ratio techniques. Jöreskog developed an early version of LISREL for confirmatory factor analysis (where latent factor relationships are correlations rather than causal; i.e., they do not have arrows). Later the method was extended to allow causality. In psychometrics and cognate fields, 'structural equation modeling' (path modeling with latent variables) is sometimes used for causal inference and sometimes to get parsimonious descriptions of covariance matrices. For causal inference, questions of stability are central. If no causal inferences are made, stability under intervention is hardly relevant; nor are underlying equations 'structural' in the econometric sense described earlier. The statistical assumptions (independence, distributions of error terms constant across subjects, parametric models for error distributions) would remain on the table. The confirmatory model testing provided by LISREL, AMOS and other programs are the primary tools of descriptive analysis for hypothesis testing and theory confirmation for complex models with latent constructs. Turner and Stevens (Turner and Stevens 1959) seminal paper introduced some of the more involved concepts of inner and outer models in path analysis structural equation modeling. Jöreskog's LISREL notation for structural equation models introduces a plethora of Greek symbols. The structural (latent variable) and measurement (indicator or measured factor) submodels are written in LISREL notation as: (1) ; (2) ; and (3) . Furthermore in order to identify a LISREL model, parameters (and more) have to be constrained by setting their values to 0 , 1 or by setting various parameters to be equal. All of this is mindboggling and Greek to most users.

Whether this notation reflects physics envy – the prejudice that anything researchable can be expressed in notation worthy of Newtonian mechanics – or merely excessive math enthusiasm, Jöreskog's Greek and index pushing does little to improve usefulness, while leaving researchers mucking through a swamp of notation. A more civilized view is embraced by the authors of rival software AMOS (McArdle and Epstein 1987; McArdle 1988) who dispense with anything but path model diagrams in their AMOS-graphics software user interface, without losing expressability or generalizability in the ensuing statistical analysis (Blunch 2008). AMOS (now an IBM product) advertises that they have dispensed with the equational minutia, and make it possible for researchers to concentrate on the path model.

There is a dedicated group of software packages which solely derived path coefficients through covariance structure modeling on Normally distributed data. These packages have the advantage that they can generate goodness-of-fit statistics for the path model as a whole, though at the expense of only being able to process Normal observations. With large enough datasets, it is argued, Central Limit Theorem convergence will allow these methods to be used for well-behaved non-Normal data. Unfortunately, as the LISREL and AMOS manuals observe, the increase in sample size may be several orders of magnitude. Software packages for covariance structure modeling have been reviewed in several survey papers (Harris 1999; Marsh, Byrne, and Craven 1992; Dhrymes et al. 1972; Dhrymes and Erlat 1972 (Marsh, Byrne, and Craven 1992; Lydtin et al. 1980))(Hox 1995). (Hox 1995) asserts that the significant contrasts appear in fit statistics, both number and applicability to specific problems. John Fox (Fox 2006) argues that many of these fit statistics are ad hoc and difficult to interpret In another option the *sem* procedure in R-language uses the reticular action model (RAM) formulation (see (Fox 2006)(Joreskog and Van Thillo 1972)) of covariance structure models. To illustrate, *sem*'s author John Fox reformulates the classic SEM model of (Blau, Duncan, and Tyree 1967) to a RAM format. Fox also implements, in his SEM software, the reticular action model (RAM) formulation of (McArdle 1988; J.P.A. Ioannidis 2005) and (McArdle and Anderson 2001). dispensing which the plethora of Greek notation that he feels overly complicates Jöreskog's formulation.

Software	Description
AMOS (IBM)	Structural equations models, multiple regression, multivariate regression, confirmatory factor analysis, structured means analysis, path analysis, and multiple population comparisons . Many consider the GUI to be the best of the commercial covariance method packages. Developed by James Arbuckle, now part of IBM through its purchase of SPSS.
CALIS (SAS)	A SAS Proc which implements multiple and multivariate linear regression; linear measurement-error models; path analysis and causal modeling; simultaneous equation models with reciprocal causation; exploratory and confirmatory factor analysis of any order; canonical correlation; a wide variety of other (non)linear latent variable models CALIS (Hartmann, 1992)
EQS (MSI.)	Structural equations models, multiple regression, multivariate regression, confirmatory factor analysis, structured means analysis, path analysis, and multiple population comparisons (Bentler, 1985, 1995). There exists an R/EQS Interface [CRAN]
LISREL (SSI)	Evolved from Karl Jöreskog's branch of algorithm development as a student of Herman Wold, LISREL (Jöreskog & Sörbom, 1989, 1996) was the first computer based covariance structure modeling package (implemented on mainframes in the late 1970s).

(continued)

Software	Description
Mplus (Muthén & Muthén)	Exploratory factor analysis; Structural equation modeling; Item response theory analysis; Growth modeling; Mixture modeling (latent class analysis); Longitudinal mixture modeling (hidden Markov, latent transition analysis, latent class growth analysis, growth mixture analysis);Survival analysis (continuous- and discrete-time);Multilevel analysis; Complex survey data analysis; Bayesian analysis; Monte Carlo simulation. There exists an R/Mplus interface automating Mplus Model Estimation and Interpretation [CRAN]
OpenMx/OpenSEM (Virginia)	A very active package that “is free and open source software for use with R that allows estimation of a wide variety of advanced multivariate statistical models contributed by experts in R and SEM. Cross platform Mac OS X
Windows XP	Windows Vista

(continued)

Software	Description
and several varieties of Linux; Open Source with Integration with R statistical language; Covariance Modeling With Means; Missing Data; Categorical Threshold Estimation; Hierarchical Model Definition; Matrix Algebra Calculationsn; User Specified Functions for Model Specification; User Specified Objective Functions; Community Wiki and Forums (Boker et al	2011) sem (R),(John Fox
2006):The first R package for SEM fit by maximum likelihood assuming multinormality	and single-equation estimation for observed-variable models by two-stage least.squares. It was also the first package I tried to run SEM in R. This implements the RAM formulation of covariance structure models (Fox
2006) TETRAD (CMU),Cross platform Mac OS X	Windows XP
Windows Vista	and several varieties of Linux; Open Source

(continued)

Software	Description
Community Forums lavaan,A promising package for SEM. Its command language is similar to those of Mplus. Hence it is perhaps the most user-friendly package for SEM to date. (Yves Rosseel	2012) semGOF,an add-on package which provides fourteen goodness-of-fit indeces for structural equation models using 'sem' package.[CRAN] SEMplusR, Functions, examples and datasets to learn, use and teach Structural Equation Modeling (GitHub) SEMModComp,Model Comparisons for SEM (CRAN) stremo,Functions to help the process of learning structural equation modelling (CRAN) semTools,Useful tools for structural equation modeling(CRAN) simsem,SIMulated Structural Equation Modeling (CRAN)

3.3 LISREL Performance Statistics

A cottage industry in ad-hoc fit indices and their evaluation have developed around covariance structure methods. It should be noted up front that a 'good fit' is not the same as strength of relationship: one could have perfect fit when all variables in the model were totally uncorrelated, as long as the researcher does not instruct the SEM software to constrain the variances. In fact, the lower the correlations stipulated in the model, the easier it is to find 'good fit.' The stronger the correlations, the more power SEM has to detect an incorrect model. When correlations are low, the researcher may lack the power to reject the model at hand. Also, all measures overestimate goodness of fit for small samples, though RMSEA and CFI are less sensitive to sample size than others (see (Cheung and Rensvold 2002)) In cases where the variables have low correlation, the structural (path) coefficients will be low also. Researchers should report not only goodness-of-fit measures but also should report the structural coefficients so that the strength of paths in the model can be assessed. Likewise, one can have good fit in a misspecified model. One indicator of this occurring is if there are high modification indexes (MI) in spite of good fit. High MI's indicate multicolinearity in the model and/or correlated error. All other things equal, a model with fewer indicators per factor will have a higher apparent fit than a model with more indicators per factor. Fit coefficients that reward parsimony are one way to adjust for this tendency.

Table 3.2: Performance metrics for covariance SEM analysis

Performance Metric	Description and Advantages
Root mean square residuals, or RMS residuals, or RMSR, or RMR	The closer the RMR to 0 for a model being tested, the better the model fit. RMR is the coefficient which results from taking the square root of the mean of the squared residuals, which are the amounts by which the sample variances and covariances differ from the corresponding estimated variances and covariances, estimated on the assumption that your model is correct. Fitted residuals result from subtracting the sample covariance matrix from the fitted or estimated covariance matrix. LISREL computes RMSR. AMOS does also, but calls it RMR.
Standardized root mean square residual, Standardized RMR (SRMR)	The smaller the standardized RMR, the better the model fit. SRMR is the average difference between the predicted and observed variances and covariances in the model, based on standardized residuals. Standardized residuals are fitted residuals (see above) divided by the standard error of the residual (this assumes a large enough sample to assume stability of the standard error). SRMR is 0 when model fit is perfect.

Model chi-square.	Model chi-square, also called discrepancy or the discrepancy function, is the most common fit test, printed by all computer programs. AMOS outputs it as CMIN. The chi-square value should not be significant if there is a good model fit, while a significant chi-square indicates lack of satisfactory model fit. That is, chi-square is a 'badness of fit' measure in that a finding of significance means the given model's covariance structure is significantly different from the observed covariance matrix. If model chi-square $< .05$, the researcher's model is rejected. LISREL refers to model chi-square simply as chi-square, but synonyms include the chi-square fit index, chi-square goodness of fit, and chi-square badness-of-fit. Model chi-square approximates for large samples what in small samples and loglinear analysis is called, the generalized likelihood ratio. There are three ways, listed below, in which the chi-square test may be misleading. Because of these reasons, many researchers who use SEM believe that with a reasonable sample size (ex., > 200) and good approximate fit as indicated by other fit tests (ex., NNFI, CFI, RMSEA, and others discussed below), the significance of the chi-square test may be discounted and that a significant chi-square is not a reason by itself to modify the model. The more complex the model, the more likely a good fit. In a just-identified model, with as many parameters as possible and still achieve a solution, there will be a perfect fit. Put another way, chi-square tests the difference between the researcher's model and a just-identified version of it, so the closer the researcher's model is to being just-identified, the more likely good fit will be found. The larger the sample size, the more likely the rejection of the model and the more likely a Type II error (rejecting something true). In very large samples, even tiny differences between the observed model and the perfect-fit model may be found significant. The chi-square fit index is also very sensitive to violations of the assumption of multivariate normality. When this assumption is known to be violated, the researcher may prefer Satorra-Bentler scaled chi-square, which adjusts model chi-square for non-normality.

Hoelter's critical N	is the size the sample size must reach for the researcher to accept the model by chi-square, at the .05 or .01 levels. This throws light on the chi-square fit index's sample size problem. Hoelter's N should be greater than 200.
Satorra-Bentler scaled chi-square	Sometimes called Bentler-Satorra chi-square, this is an adjustment to chi-square which penalizes chi-square for the amount of kurtosis in the data. That is, it is an adjusted chi-square statistic which attempts to correct for the bias introduced when data are markedly non-normal in distribution.
Relative chi-square, also called normal chi-square	is the chi-square fit index divided by degrees of freedom, in an attempt to make it less dependent on sample size. (Carmines & McIver, 1981) states that relative chi-square should be in the 2:1 or 3:1 range for an acceptable model. Some researchers allow values as high as 5 to consider a model adequate fit, while others insist relative chi-square be 2 or less. AMOS lists relative chi-square as CMIN/DF.
Goodness-of-fit index, GFI (Jöreskog-Sörbom GFI)	GFI varies from 0 to 1, but theoretically can yield meaningless negative values. A large sample size pushes GFI up. Though analogies are made to R-square, GFI cannot be interpreted as percent of error explained by the model. Rather it is the percent of observed covariances explained by the covariances implied by the model. That is, R2 in multiple regression deals with error variance whereas GFI deals with error in reproducing the variance-covariance matrix. As GFI often runs high compared to other fit models, some suggest using .95 as the cutoff. By convention, GFI should by equal to or greater than .90 to accept the model. LISREL and AMOS both compute GFI.

Adjusted goodness-of-fit index, AGFI	AGFI is a variant of GFI which adjusts GFI for degrees of freedom: the quantity (1 - GFI) is multiplied by the ratio of your model's df divided by df for the baseline model, then AGFI is 1 minus this result. AGFI can yield meaningless negative values. AGFI $>$ 1.0 is associated with just-identified models and models with almost perfect fit. AGFI $<$ 0 is associated with models with extremely poor fit, or based on small sample size. AGFI should also be at least .90. Like GFI, AGFI is also biased downward when degrees of freedom are large relative to sample size, except when the number of parameters is very large. Like GFI, AGFI tends to be larger as sample size increases; correspondingly, AGFI may underestimate fit for small sample sizes, according to (Bollen, 1989). The goodness-of-fit index (GFI) and the adjusted goodness-of-fit index (AGFI) are ad hoc measures of the descriptive adequacy of the model. Although the GFI and AGFI are thought of as proportions, comparing the value of the fitting criterion for the model with the value of the fitting criterion when no model is fit to the data, these indices are not constrained to the interval 0 to 1. Several rough cutoffs for the GFI and AGFI have been proposed; a general theme is that they should be close to 1. It is probably fair to say that the GFI and AGFI are of little practical value.
Centrality index, CI	CI is a function of model chi-square, degrees of freedom in the model, and sample size. By convention, CI should be .90 or higher to accept the model.
Noncentrality parameter, NCP	This is also called the McDonald noncentrality parameter index and DK, is chi-square penalizing for model complexity. To force it to scale to 1, the conversion is exp(-DK/2). NCP is used with a table of the noncentral chi-square distribution to assess power. RMSEA, CFI, RNI, and CI are related to the noncentrality parameter. (Raykov, 2005) has argued that fit measures based on noncentrality are biased.

Goodness-of-fit tests comparing the given model with an alternative model	This set of goodness of fit measures compare your model to the fit of another model. This is well and good if there is a second model. When none is specified, statistical packages usually default to comparing your model with the independence model, or even allow this as the only option. The independence model is the null model, which is the model in which variables are assumed to be uncorrelated with the dependent(s). Since the fit of the independence model is usually terrible, comparing your model to it will generally make your model look good but may not serve your research purposes.
The comparative fit index, CFI	CFI is also known as the Bentler Comparative Fit Index. CFI compares the existing model fit with a null model which assumes the latent variables in the model are uncorrelated (the ’independence model’). That is, it compares the covariance matrix predicted by the model to the observed covariance matrix, and compares the null model (covariance matrix of 0’s) with the observed covariance matrix, to gauge the percent of lack of fit which is accounted for by going from the null model to the researcher’s SEM model. Note that to the extent that the observed covariance matrix has entries approaching 0’s, there will be no non-zero correlation to explain and CFI loses its relevance. CFI is similar in meaning to NFI (see below) but penalizes for sample size. CFI and RMSEA are among the measures least affected by sample size (Fan et al., 1999). CFI varies from 0 to 1. CFI close to 1 indicates a very good fit. CFI is also used in testing modifier variables (those which create a heteroscedastic relation between an independent and a dependent, such that the relationship varies by class of the modifier). By convention, CFI should be equal to or greater than .90 to accept the model, indicating that 90% of the covariation in the data can be reproduced by the given model.
GFI based on predicted vs. observed covariances, penalizing lack of parsimony	Parsimony measures. These measures penalize for lack of parsimony, since more complex models will, all other things equal, generate better fit than less complex ones. They do not use the same cutoffs as their counterparts (ex., PCFI does not use the same cutoff as CFI) but rather will be noticeably lower in most cases. Used when comparing models, the higher parsimony measure represents the better fit.

Parsimony ratio (PRATIO)	PRATO is the ratio of the degrees of freedom in your model to degrees of freedom in the independence (null) model. PRATIO is not a goodness-of-fit test itself, but is used in goodness-of-fit measures like PNFI and PCFI which reward parsimonious models (models with relatively few parameters to estimate in relation to the number of variables and relationships in the model). See also the parsimony index, below.
Parsimony index	The parsimony index is the parsimony ratio times BBI, the Bentler/Bonnett index, discussed above. It should be greater than .9 to assume good fit.
Root mean square error of approximation	RMSEA, is also called RMS or RMSE or discrepancy per degree of freedom. By convention, there is good model fit if RMSEA less than or equal to .05. There is adequate fit if RMSEA is less than or equal to .08. More recently, (Hu & Bentler, 1999) have suggested RMSEA <= .06 as the cutoff for a good model fit. RMSEA is a popular measure of fit, partly because it does not require comparison with a null model and thus does not require the author posit as plausible a model in which there is complete independence of the latent variables as does, for instance, CFI. Also, RMSEA has a known distribution, related to the non-central chi-square distribution, and thus does not require bootstrapping to establish confidence intervals. Confidence intervals for RMSEA are reported by some statistical packages. It is one of the fit indexes less affected by sample size, though for smallest sample sizes it overestimates goodness of fit (Fan et al., 1999).
Goodness of fit measures based on information theory	Measures in this set are appropriate when comparing models which have been estimated using maximum likelihood estimation. As a group, this set of measures is less common in the literature, but that is changing. All are computed by AMOS. They do not have cutoffs like .90 or .95. Rather they are used in comparing models, with the lower value representing the better fit.

Chapter 4

Systems of Regression Equations

4.1 The Birth of Structural Equation Modeling

Alfred Cowles III hailed from an established Chicago publishing family, his father and uncle having founded the Chicago Tribune and Cleveland Leader respectively (Grier 2013). For a short time after WWI Cowles successfully ran a Chicago investment firm that acquired and restructured small railroads. His firm also published a stock market newsletter providing fundamental analysis and recommendations on railroad stock issues as well as other investments, and for a time there was even an Alfred Cowles Railroad.

Diagnosed with tuberculosis in the late 1920s, Cowles consolidated his investments (just prior to the 1929 crash) and moved to Colorado Springs in search of better health (Grier 2013). Consigned to a life of enforced leisure, he filled his time developing linear regression models that simultaneously compared the predictions of 24 stock market newsletters to actual stock performance. Cowles quickly came to the conclusion that forecasters were guessing; that they offered little useful investment information, and were more often wrong than right (Cowles, 1933). Understandably, he also applied his regression skills to investigate whether good climates, like Colorado Springs, improved the outcome of tuberculosis (Cowles 3rd and Chapman 1935) with somewhat more optimistic results.

The pen and paper calculation required at the time for the regression formulas he used soon exceed his capabilities as a lone researcher. At this point he made a decision to invest some of his fortune to create the Cowles Commission, an institution dedicated to linking economic theory to mathematics and statistics. To that end, its mission was to develop a specific, probabilistic framework for estimating simultaneous regression equations to model the U.S. economy.

The Cowles Commission moved from Colorado Springs to the University of Chicago in 1939 where economist Tjalling Koopmans (Koopmans and Beckmann 1957) developed the systems of regression tools that Cowles originally had sought. This period also expanded Cowles personal files into what ultimately became the Compustat and CRSP databases, and created the market index that eventually became the Standard & Poor's 500 Index. Throughout its 15 years at the University of Chicago, the Commission clashed repeatedly with the Economics

Department and in 1955, ultimately made the decision to move to Cowles' alma mater Yale, where it was renamed the Cowles Foundation.

The Cowles Commission's most important contribution to statistics was in exposing the bias of ordinary least squares regression coefficient estimates. Cowles researchers developed new methods such as the indirect least squares, instrumental variable methods, full information maximum likelihood method, and limited information maximum likelihood methods to resolve this problem (Christ 1994).

Eleven Cowles associates ultimately received the Nobel Prize in Economics, most notably (for this book) Trygve Haavelmo, who introduced his Scandinavian colleagues Herman Wold, and Karl Jöreskog to Cowles' simultaneous regression equation approaches. Wold ultimately went on to develop PLS-PA and Jöreskog developed LISREL as latent variable alternatives that they considered more suitable for the abstract and unstructured problems of sociology, education and psychology.

4.2 Simultaneous Regression equation models

While Wold and Joreskog were pursuing idiosyncratic solutions to path coefficients, work at the Cowles Commission, under Koopmans, Zellner, Anderson, Dhrymes, and many others, made rapid progress in devising econometric tools for the networked relationships found in the U.S. economy. Their simultaneous equations regression (also 'systems of regression equation') approaches now comprise the mainstream approach in econometrics and other fields for mapping network relationships between variables. In general, this line of research has eschewed working with latent variables, but only because there was no need for special methods for dealing with them – they are linear functions of indicators, with coefficients set by the first principal component. We will explore the use of systems of regression equation approaches with latent variables later in this chapter. Simultaneous equation models are a multi-equation regression model in the form of a set of linear simultaneous equations, where the covariance matrix is not diagonal (i.e. there is covariance between the separate linear equations). It is extremely common in econometrics to encounter systems of regression equations (which need to be estimated simultaneously, i.e., as a network of relationships across equations). The equations are written in vector-matrix form, and all endogenous variables are algebraically moved to the left-hand side to produce what is called the 'structural form' system of equations. It was this usage of structural form that was adopted by Wold and Jöreskog to describe their particular setups, which is why they called them structural equation models (SEM). Further algebraic manipulation to pull all endogenous variables to the left-hand side and exogenous variables to the right-hand side of the equation produce the 'reduced form' system of equations. The reduced form is a simple general linear model which may be estimated using ordinary least squares regression.

Unfortunately, the task of decomposing the estimated matrix algebraically into the individual factors is often complicated. There may indeed be questions concerning whether the estimators for the original equation can be algebraically recovered and are unique – it may be possible to have no solutions, or to have an infinite number of solutions derived from the reduced form. To assure that the estimators we recover from the reduced form are unique, we apply specific identification conditions before estimating. If these are not met a model restructuring is required.

In order for a unique estimate to be derived, three conditions must be met:

1. the error terms are assumed to be serially independent and identically distributed
2. the rank of the matrix of exogenous regressors must be equal to the number of exogenous regressors.
3. the identification conditions requires that the number of unknowns in this system of equations not exceed the number of equations. There are two identification conditions:
   a. the order condition requires that the number of excluded exogenous variables is greater or equal to the number of included endogenous variables; and
   b. the rank condition states puts constraints on the rank of the matrix which is obtained from the reduced form exogenous coefficient matrix by crossing out those columns which correspond to the excluded endogenous variables, and those rows which correspond to the included exogenous variables.

Path analysis using linear simultaneous equations have the following advantages:

1. They describe path coefficients in terms of regression coefficients (Tukey, 1954 claimed they were more informative than correlations; and easier to interpret than covariances)
2. They are full information methods (versus PLS path analysis which is limited information)
3. They have well defined performance metrics (fit statistics) and analysis approaches, including residual analysis for underlying model assumptions; neither PLS path analyses nor covariance methods have this). In particular, hypothesis tests are well-defined, and can convincingly reject alternative hypotheses.
4. They allow for residuals that can be plotted and inspected for data problems such as autocorrelation, heteroskedacicity, non-Normality, outliers and more. The two other approaches do not allow this
   a. PLS path analysis obscures any analysis of this sort because of resampling; and
   b. iterative search algorithms that underlie covariance solutions obscure the impact of non-Normal and problem data on residuals.
5. There are transformations that are well understood (logit, probit, log, Box-Cox, etc.) that can be used on non-linear data

Consider an example of a reformulation of a latent variable structural model in a fashion that allows systems of equation estimation of a path model 4.1.

The reduced form of this system can be estimated with OLS regression, and the original parameters (and thus path coefficients on the structural model) can be recovered by reversing the algebraic transformation that yielded the reduced form. The standard identification conditions (rank and order) apply for identification (Greene & Zhang, 2003). When a model

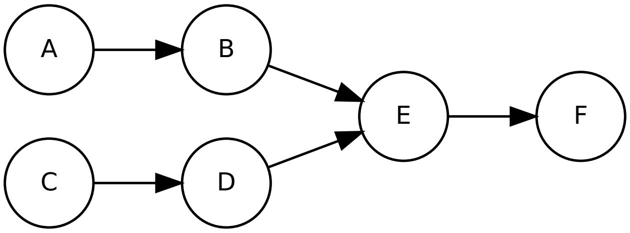

Figure 4.1: A latent variable (inner) path model

is identified, this means that unique estimates for the path coefficients can be obtained; over or under-identification results in either multiple estimates for each path or none at all.

4.3 Estimation

The most common estimation method for the simultaneous equations models are:

1. The two-stage least squares method, developed independently by (Theil and Boot 1962) and (Basmann 1963b). It is an equation-by-equation technique, where the endogenous regressors on the right-hand side of each equation are being instrumented with the regressors from all other equations.

2. The indirect least squares is an approach in econometrics where the coefficients in a simultaneous equations model are estimated from the reduced form model using ordinary least squares. For this, the structural system of equations is transformed into the reduced form first. Once the coefficients are estimated the model is put back into the structural form.

3. The "limited information" maximum likelihood method was suggested in (Anderson, Rubin, and others 1950) (TW Anderson 2005b).

4. The three-stage least squares estimator was introduced by (Arnold Zellner and Theil 1962). It combines two-stage least squares (2SLS) with seemingly unrelated regressions (SUR). There are variations on the method, including i-3SLS which involves an iterative search for estimators.

5. A seemingly unrelated regression (SUR) estimation procedure may be used f the error terms are not independent. Seemingly unrelated regressions consist of several regression equations, each having its own dependent variable and potentially different sets of exogenous explanatory variables. Equation by equation estimates are consistent, however generally not as efficient as the SUR method. When the covariance matrix is known to be diagonal, that is, there are no cross-equation correlations between the error terms. In this case the system becomes not seemingly, but truly unrelated.

Table 4.1: Systems of Equations Regression Software

Software	Features
systemfit	R's systemfit package can estimate systems of linear equations within the R programming environment and can be used for ordinary least squares OLS; seemingly unrelated regression SUR; and the instrumental variable IV methods; two-stage least squares 2SLS and three-stage least squares 3SLS where SUR and 3SLS estimations can optionally be iterated. The systemfit package provides tools for several statistical tests. It has been tested on a variety of datasets and its reliability is demonstrated
SAS	PROC MODEL estimates ARIMA PDL dynamic modeling supports the following methods for parameter estimation (1) ordinary least squares OLS (2) two-stage least squares 2SLS (3) seemingly unrelated regression SUR and iterative SUR ITSUR (4) three-stage least squares 3SLS and iterative 3SLS IT3SLS (5) generalized method of moments GMM (6) simulated method of moments SMM (7) full information maximum likelihood FIML (8) general log-likelihood maximization (9) simulation and forecasting capabilities (9) Monte Carlo simulation and (10) goal-seeking solutions
STATA	STATA's reg3 Command estimates OLS 2SLS and 3SLS with some limitations
SPSS/Systat/AMOS	Neither SPSS nor Systat packages support estimation of 3SLS or FIML AMOS package estimates 2SLS
Eviews	Windows-based econometric and forecasting software has object-oriented interface to powerful statistical forecasting and modeling tools
LIMDEP	Single-equation and simultaneous-equation regression models
MATLAB/Octave / Gauss and Excel	Computational software that is sometimes redeployed for simultaneous equation regression analysis

Table 4.2: Performance Metrics for Systems of Equations Regression

Performance.metric	Application
R-squared	In statistics the coefficient of determination R-squared is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information It is the proportion of variability in a data set that is accounted for by the statistical model It provides a measure of how well future outcomes are likely to be predicted by the model There are several different definitions of R-squared which are only sometimes equivalent One class of such cases includes that of linear regression In this case if an intercept is included then R-squared is simply the square of the sample correlation coefficient between the outcomes and their predicted values or in the case of simple linear regression between the outcomes and the values of the single regressor being used for prediction In such cases the coefficient of determination ranges from 0 to 1 Important cases where the computational definition of R-squared can yield negative values depending on the definition used arise where the predictions which are being compared to the corresponding outcomes have not been derived from a model-fitting procedure using those data and where linear regression is conducted without including an intercept Additionally negative values of R-squared may occur when fitting non-linear trends to data In these instances the mean of the data provides a fit to the data that is superior to that of the trend under this goodness of fit analysis
F-test	An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis It is most often used when comparing statistical models that have been fit to a data set in order to identify the model that best fits the population from which the data were sampled Exact F-tests mainly arise when the models have been fit to the data using least squares
t-statistics	on individual parametersThe t-statistic is a ratio of the departure of an estimated parameter from its notional value and its standard error It is used in hypothesis testing for example in the Students t-test in the augmented DickeyFuller test
Graphical examination of plots of regression residuals	Flexible ad hoc method of testing model assumptions The following four assumptions on the random errors are equivalent to the assumptions on the response variables; i. The random errors are independent; ii. The random errors are normally distributed; iii. The random errors have constant variance; iv. The random errors have zero mean

Table 4.2: Performance Metrics for Systems of Equations Regression *(continued)*

Performance.metric	Application
Miscellaneous other test statistics	Breusch-Godfrey test; Breusch-Pagan test; Cook's distance; DFFITS; Goldfeld-Quandt test; Leverage; Park test; Partial leverage; Partial regression plot; Partial residual plot; Portmanteau test; PRESS statistic; Ramsey RESET test; Regression model validation; Variance inflation factor; White test

4.4 Comparing the Different SEM Methods

Researchers often ask whether estimators computed by different SEM approaches and software are comparable and consistent. Certainly, commercial software for PLS path analysis and covariance structure methods can be inconsistent in its application of assumptions, algorithms and reporting standards. In practice, if we can extract sufficient model-specific information from the data, the various methods will yield comparable, and often identical estimates The caveat here is whether or not there is sufficient information in the dataset, which in turn depends on sample size, soundness of research design and informativeness of the measurements. As we have seen previously, PLS-PA will allow inference with insufficient sample sizes from poorly designed instruments, with the consequent inaccuracies in estimation. A comparison of methodologies needs to be standardized in two sets of specific computations – (1) the computation of factor weights (on the so-called 'outer' model), which determine realized values of the associated latent variable) and (2) the reported path coefficients between latent variables (the so-called 'inner' model). Path analysis software usually allows pre-selection of the factors that comprise particular latent variable (i.e., reflective links). The weights assigned are most often the factor weights of the first principal component. This is the approach applied here. Reported path coefficients between latent variables can be:

- 0-dimensional (correlations),
- 1-dimensional (regression coefficients) or
- 2-dimensional (covariances).

Wright's original path analysis reported correlations (dimensionless), which Wright argued were easy to interpret, and less likely lead to erroneous conclusions. These problems are less significant when all measurements are encoded on the same Likert-scale, because no matter what the dimensionality of the path coefficients, the estimators will look like correlations. The reason is that Likert-scales are essentially dimensionless, intended to capture unobservable human utilities without reference to a tangible standard. Likert measurements are scaled relative to other Likert measurements in a particular study context, and this makes them dimensionless. Tukey promoted regression coefficients (1-dimensional) on the path because they provide information on scale as well as strength of the link. Such path coefficients are typically computed through a sequence of pairwise latent variable OLS regressions (following (Lohmoller, 1988; Lohmöller, 1989). Some software allow alternatives, including PLSR,

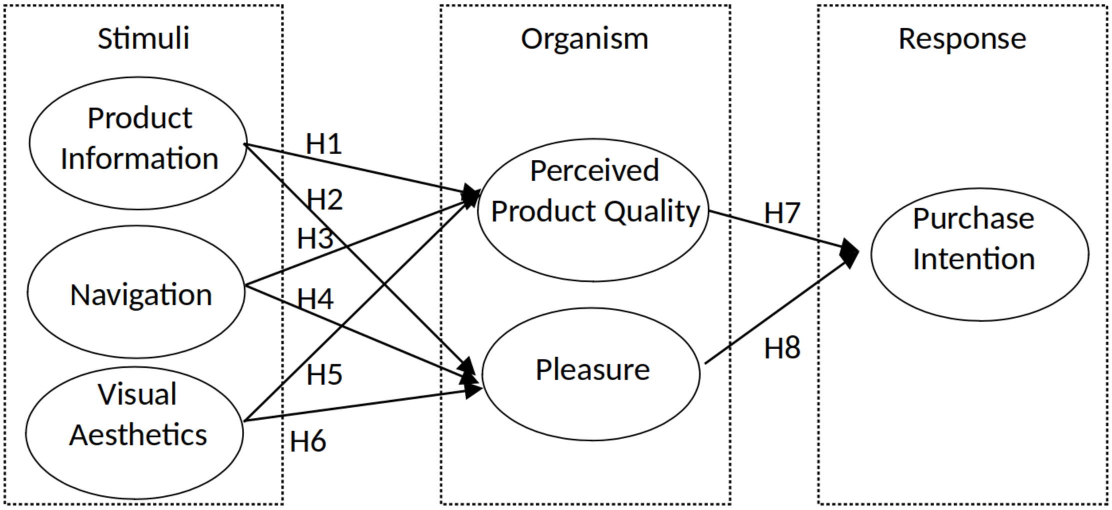

Figure 4.2: (#fig:theory_model)Conceptual 'theory' assumed to generate observations in the dataset

though with little difference in results. PLSR provides new insights only where the study involves a large, multicolinear dataset, as in spectroscopy, chemometrics, and some other natural sciences. For most practical purposes, regression coefficients will be identical for OLS, PCR and PLSR applied piecewise to path regression, and what differences exist are overwhelmed by the effects of resampling. Covariance methods naturally generate covariances (2-dimensional) though these are difficult to interpret, and are overly sensitive to the scale of interaction. Because of this, software packages like LISREL and AMOS standardize path coefficients to dimensionless correlations (following Wright) or offer alternatives that compute regression coefficients. In a typical study, for example, the first step in SEM path analysis would be to choose latent variables, and the construction of a structural model relating them. Usually this starts with a theory, hypothesis or hunch about the way a process works in the real world population that is of interest. For our comparison of methods, we conducted a study of 504 college sophomores with an average age of 19. There were an equal number of males and female students involved in this well-designed SEM analysis, including comparisons of the outcomes of PLS-PA, Covariance-based SEM, and Regression-based SEM. All of the competing or complementary statistical methodologies are applied to a single dataset of 504 7-point Likert scaled survey responses on 33 variables which have been clustered into 6 latent constructs that reflect the researchers theoretical perspective (Ye and Shiau 2010).

Structural models are often applied in survey research where the underlying theory is based on relationships between unobserved (latent) constructs. The structural model that was fitted in this section is graphically depicted in figure 4.8 and is similar to commonly used models in the social sciences. The research data considered in this section obtained data on several indicators of website quality as perceived by users, and how that data influenced their purchase intentions. Summary statistics appear in @ref(fig:sum_purch) where the overall Cronbach alpha for the dataset indicator variables is 0.92 indicating that all of the variables are essentially measuring one construct. @ref(fig:sum_purch) column headings include:

- n: number of complete cases for the item
- raw.r: The correlation of each item with the total score, not corrected for item overlap
- std.r: The correlation of each item with the total score (not corrected for item overlap) assuming the items were all standardized
- r.cor: item whole correlation corrected for item overlap and scale reliability
- r.drop: item whole correlation for this item against the scale without this item
- mean: mean of each item
- sd: standard deviation of each item

	n	raw.r	std.r	r.cor	r.drop	mean	sd
nav1	504	0.50779	0.50939	0.48845	0.45155	5.5833	1.2018
nav2	504	0.58257	0.57671	0.55111	0.53131	4.9683	1.2190
nav3	504	0.58355	0.58474	0.56733	0.53618	5.5020	1.1333
nav4	504	0.64726	0.64301	0.62841	0.60374	5.1369	1.1708
aes1	504	0.60190	0.59790	0.58454	0.55592	5.0417	1.1351
aes2	504	0.61479	0.61142	0.59942	0.56618	4.9127	1.2228
aes3	504	0.63899	0.63398	0.61720	0.59402	4.9921	1.1881
prod1	504	0.61021	0.59754	0.58041	0.55901	5.0000	1.2737
prod2	504	0.59623	0.57789	0.56359	0.54049	4.7560	1.3473
prod3	504	0.57305	0.55449	0.53665	0.51478	4.4702	1.3550
prod4	504	0.61317	0.60211	0.58155	0.56582	4.9960	1.1897
prod5	504	0.59074	0.58001	0.55240	0.53815	4.8254	1.2648
qual1	504	0.59659	0.59423	0.57874	0.54799	4.4226	1.1856
qual2	504	0.61284	0.60620	0.59680	0.56577	4.3671	1.1823
qual3	504	0.56846	0.56238	0.54645	0.51579	4.2103	1.2241
pleas1	504	0.56039	0.57606	0.55967	0.51191	4.6806	1.1188
pleas2	504	0.61307	0.62701	0.61457	0.57028	4.7262	1.0814
pleas3	504	0.53855	0.55251	0.53113	0.49108	4.6944	1.0635
pleas4	504	0.56583	0.58165	0.57213	0.51817	4.7222	1.1094
pleas5	504	0.58399	0.59871	0.58757	0.54040	4.7500	1.0480
purch1	504	0.60881	0.62122	0.61365	0.56305	4.9762	1.1435
purch2	504	0.64996	0.66202	0.65856	0.61037	5.0159	1.0774
purch3	504	0.63161	0.64281	0.63358	0.59073	5.1706	1.0713
purch4	504	0.57564	0.58012	0.55331	0.52660	4.8056	1.1567

We first calculate the sample size required for discriminating at an effect size of 0.1 (we show how to compute this in Chapter 6). An effect size of 0.1 implies that we can identify canonical correlations in excess of 0.1 on any link in the model; 0.1 is a reasonable critical factor, as correlations below this value are for practical purposes, random. In addition we have to define the power and significance of our tests. We will set these at the 0.05 significance level established by Ronald Fisher (Fisher 2006), and power level of 0.8 established by Jacob Cohen (Cohen 1988). In the absence of other context specific information, these are widely accepted

Table 4.4: PC Regression Results

	Coefficient
(Intercept)	0.39310
PC1	0.17763
PC2	0.10889
PC3	0.15113
PC4	-0.00847
PC5	0.07882
PC6	-0.07579

values for these critical factors. The model has six latent variables (product information, navigation, aesthetics, product quality, pleasure and purchase approach) and 33 indicators from a survey of 504 subjects. Our required sample size is 1,713 for such a model assuming that data has a Gaussian distribution (see (Westland 2010)) or around four times the size of our actual sample. Ideally we should collect more data to discriminate an effect of 0.1; our sample of 504 will just be able to discriminate between effects of 0.2 or more when data is Gaussian; of course, since the data is categorical Likert-scale data, a Gaussian assumption will be incorrect, and we need to expand sample size even more for unequivocal results. At our current sample size of 504 subjects, SEM models tend to overfit datasets. This is less of a problem in covariance methods, since datasets that are too small, or which are multicolinear, cause the algorithm to generate singular matrices. PLS-PA has no constraints on generating estimates, since it is only computing pairwise 'correlations', and can yield link estimators that are wildly inaccurate. Nonetheless, PLS-PA converges relatively quickly, and we see that in this example both PLS-PA and covariance methods yield similar estimates, though there is no guarantee that our estimators will not significantly overfit the data. Only larger sample sizes will guarantee reliable estimates.

The appropriate starting point for data-centric research is exploratory analysis of the data. Likert scaled data tends to hold fewer surprises for researchers as the data is discrete and bounded between zero and a relatively small integer. We are most interested in the measurements of latent constructs (as opposed to demographic qualifiers, which are present to identify biases in the experimental design). We use a principal components analysis to identify naturally occurring clusters in our subjects' perceptions. These should be reflected in our theory, and if they are not, we are obligated to revisit the theory. We may obtain a general sense of the influence of key differentiators of gender, income, purchase engagement and web engagement, by looking at their clustering (or lack of) on the principal components. The first two principal components explain over 40% of variance in this datasets, and we peruse these key differentiators projected onto these components in figures 4.3 , 4.4 , 4.5 and 4.6. There is no discernable pattern which suggests that responses are systematically biased by these demographic differentiators, and we can with confidence proceed to consider the model tested.

The rug plot (figure 4.7) of the distributions of the four "purchase approach"" indicator variables shows a consistent clustering around Likert scale responses 5 and 6 for all four

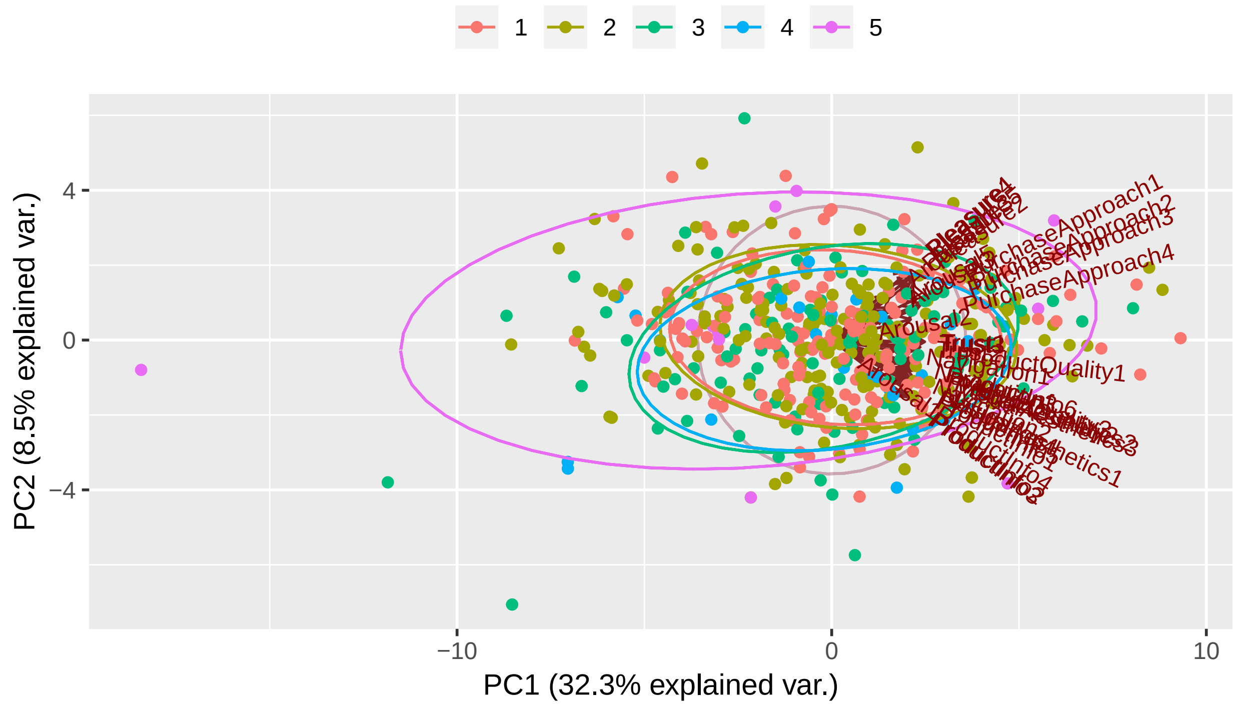

Figure 4.3: Mapping Web Engagement on the first Two Principal Components (1: < 1 hour / wk to 5: > 20 hours / wk)

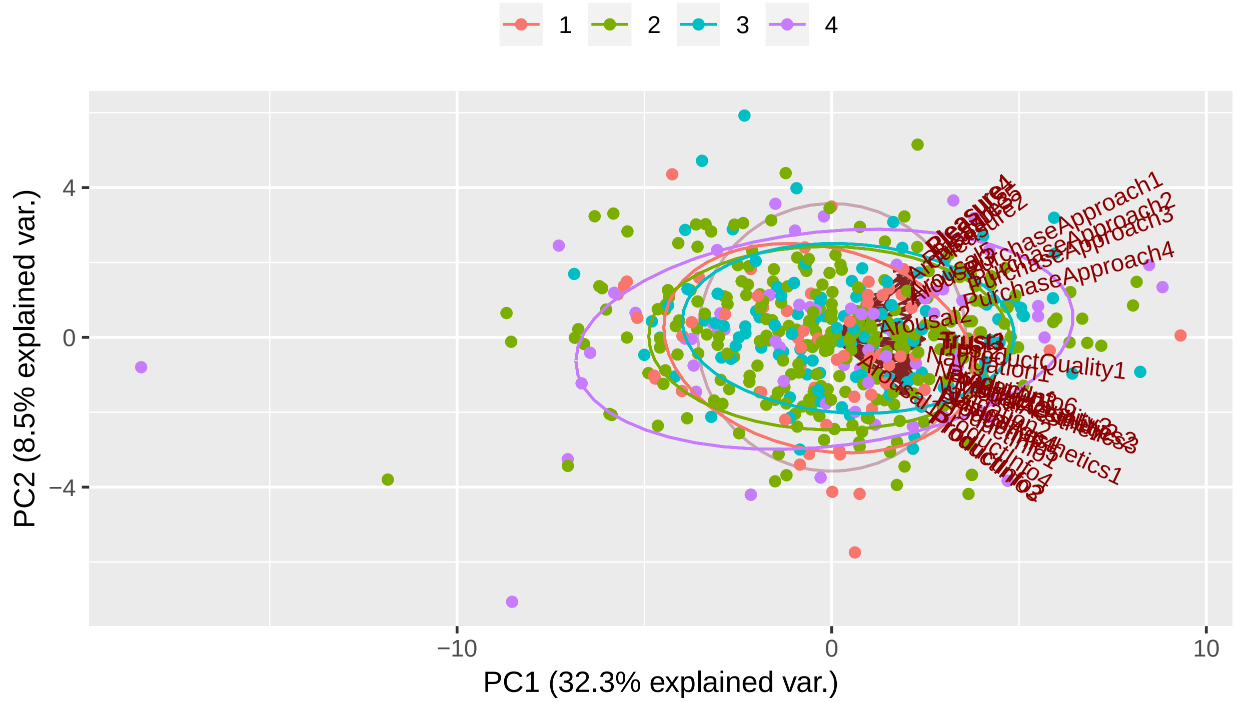

Figure 4.4: Mapping Product Engagement on the first Two Principal Components (1: < 1 hour / wk to 5: > 20 hours / wk)

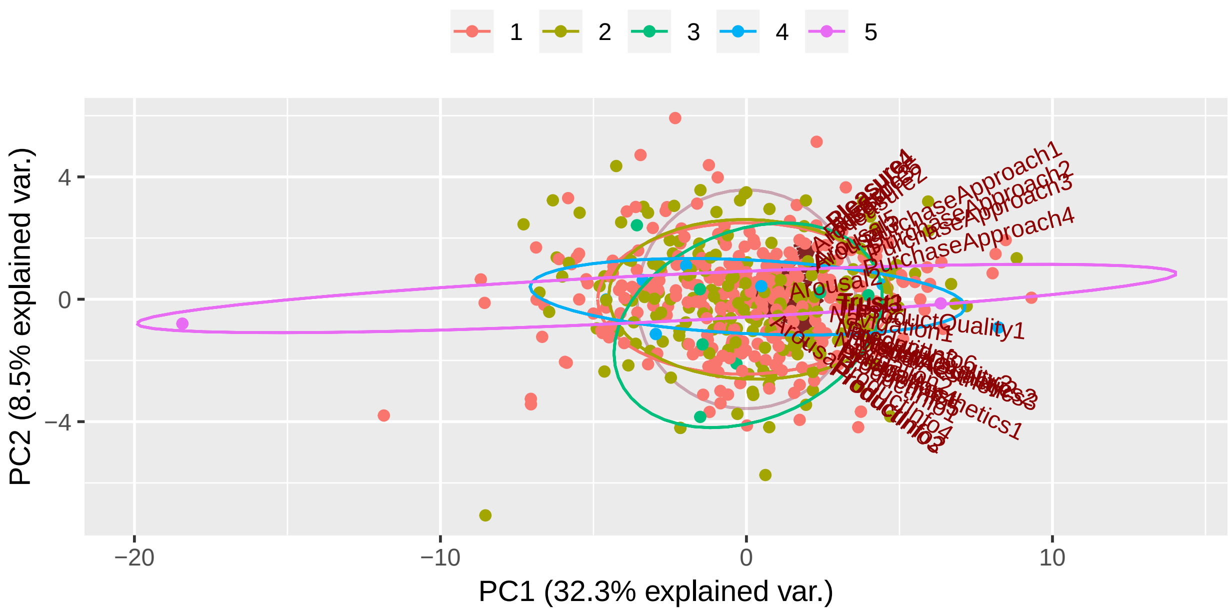

Figure 4.5: Mapping Income on the first Two Principal Components (1: < $10,000 to 5: > $100,000)

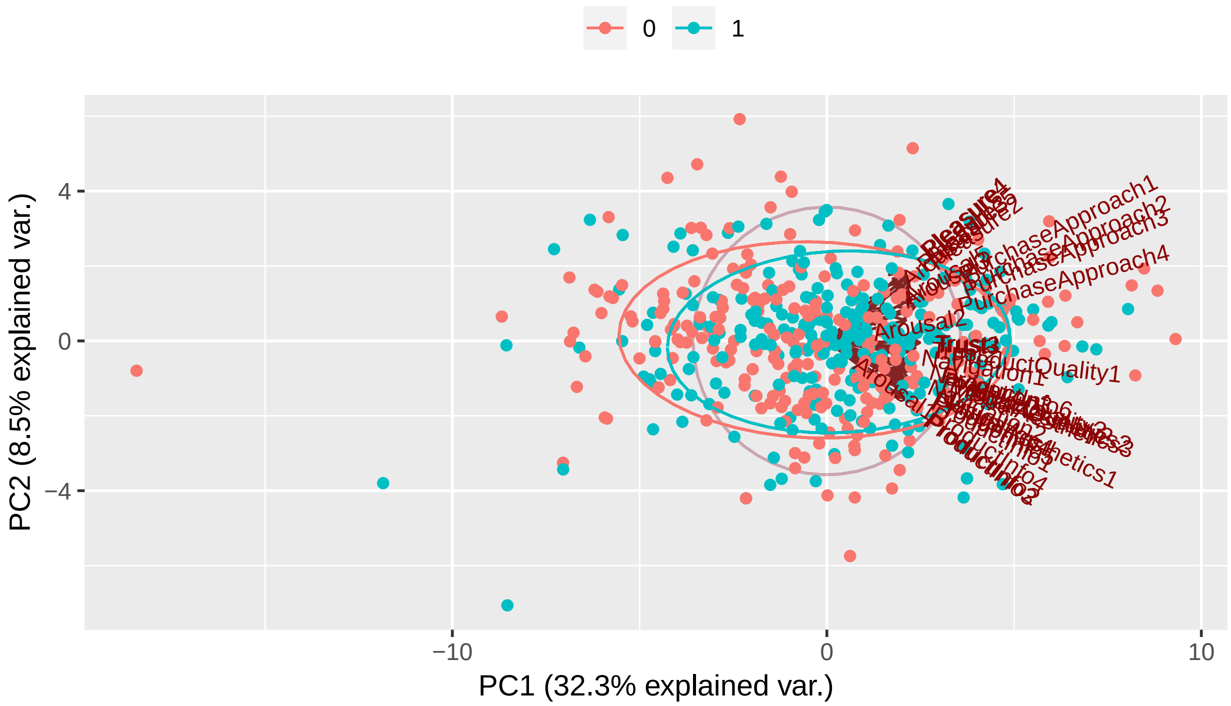

Figure 4.6: Mapping Gender on the first Two Principal Components (0=Male 1= Female)

Table 4.5: PLS-Regression Coefficients (NIPALS) for Each of the Dependent Variable Components

	PurchaseApproach1	PurchaseApproach2	PurchaseApproach3	PurchaseApproach4
Navigation1	0.16225	0.13697	0.11138	0.03579
Navigation2	-0.01129	-0.00476	0.00095	0.02739
Navigation3	0.15915	0.13661	0.11150	0.03839
Navigation4	0.01570	0.02000	0.02060	0.03128
VisualAesthetics1	0.00813	0.01435	0.01583	0.02887
VisualAesthetics2	-0.01109	-0.00267	0.00254	0.02754
VisualAesthetics3	-0.03390	-0.02156	-0.01163	0.02959
ProductInfo1	-0.02523	-0.01434	-0.00688	0.02490
ProductInfo2	-0.03958	-0.02499	-0.01615	0.01879
ProductInfo3	-0.04664	-0.03036	-0.02049	0.01757
ProductInfo4	0.05238	0.05153	0.04418	0.02760
ProductInfo5	-0.02012	-0.00926	-0.00312	0.02421
ProductInfo6	0.12509	0.11186	0.09267	0.03990
Trust1	0.17316	0.14996	0.12276	0.04442
Trust2	0.04050	0.03941	0.03664	0.03765
Trust3	0.11838	0.10659	0.08916	0.04311
ProductQuality1	0.05310	0.05167	0.04531	0.03350
ProductQuality2	-0.09849	-0.07461	-0.05320	0.02503
ProductQuality3	-0.03121	-0.01893	-0.01050	0.02443
Pleasure1	0.12156	0.10203	0.08638	0.04647
Pleasure2	0.09666	0.08290	0.07184	0.04757
Pleasure3	0.01293	0.01242	0.01630	0.03955
Pleasure4	0.01710	0.01559	0.01883	0.04009
Pleasure5	0.01611	0.01525	0.01933	0.04455
Arousal1	-0.00771	-0.00233	-0.00184	0.00023
Arousal2	-0.03703	-0.02710	-0.01984	0.00621
Arousal3	0.08465	0.07528	0.06282	0.02949
Arousal4	0.05150	0.04625	0.04110	0.03273
Arousal5	0.01205	0.01380	0.01480	0.02490
INTERCEPT	0.08183	0.20848	0.75359	0.51125

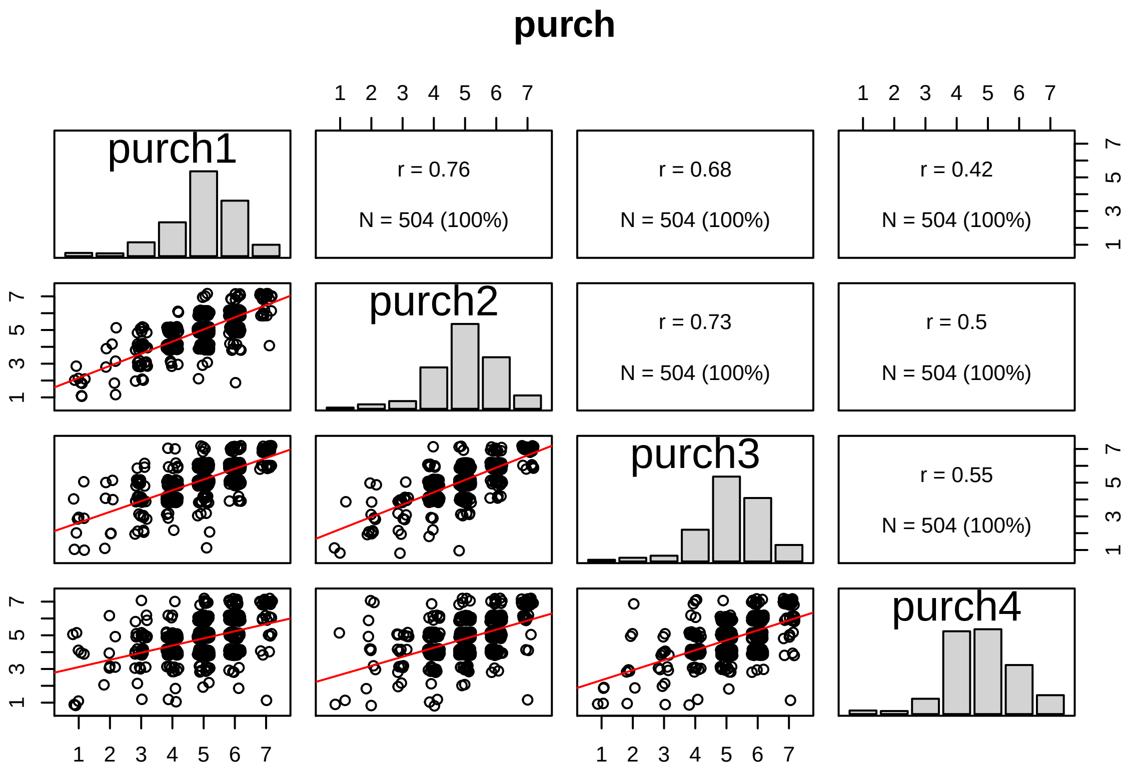

Figure 4.7: Rugplot of correlations of the four Likert scaled (7-point) 'purchase approach' indicator variable distributions (dependent variables in the PLS Regression)

indicators. This clustering suggests that the indicators (essentially single questions on the survey instrument) were appropriately selected and are measuring the same underlying latent constructs. Researchers will often use Cronbach's alpha as a measure of this clustering around latent constructs. This particular dataset was well structured in the sense that indicators consistently measured their intended latent construct.

We can compare the PLS results to a covariance based analysis (in this case the RAM-AMOS formulation (Fox 2002). Since the two methods are not equivalent, but at a macro level use comparable path models, we can make a few assumptions to convert the PLS-PA problem setup to an almost equivalent RAM-AMOS setup. For the scaling, we fix the latent variables' variances and the first loadings of each latent variable's instrument to one, and variances for the measured variables are fixed to one as well. The resulting factor loadings will not be equivalent, but this approach allows us to focus on estimation of path coefficients. We do something similar an investigating various systems of regression approaches later on in this section.

Results are relatively consistent across all the methods, though some of this consistency was due to purposely setting up the assumptions, methods and data representation to make results comparable, and all of the data uses the same Likert scale. I have also used two popular commercial packages, SmartPLS and AMOS, to analyze the same data for comparison.

Bearing in mind that the sample size was too small for the models estimated, the results are

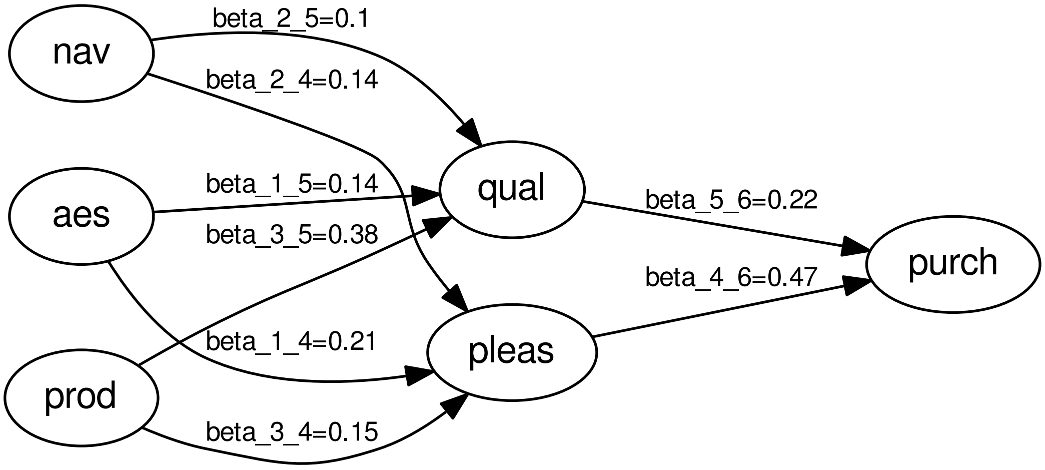

Figure 4.8: PLS-PA Graph (output from package semPLS)

Table 4.6: Summary of AMOS Path Coefficients

Path	Estimate	CBSEM
aes -> pleas	0.21362	0.05641
nav -> pleas	0.14019	0.37385
prod -> pleas	0.14803	-0.07708
aes -> qual	0.13732	0.25382
nav -> qual	0.10335	-0.02449
prod -> qual	0.38221	0.29443
pleas -> purch	0.46856	0.10976
qual -> purch	0.21561	0.30788

Table 4.7: Summary of Path Coefficients from All Methods

Path	sem	AMOS	semPLS	SmartPLS	OLS	WLS	SUR	TwoSLS	ThreeSLS
aes -> pleas	0.21	0.27	0.21	0.21	0.38	0.38	-0.23	2.43	0.78
nav -> pleas	0.14	0.24	0.14	0.17	0.39	0.39	-0.3	1.86	2.33
prod -> pleas	0.15	0.16	0.15	0.13	0.49	0.45	0.97	1.58	-6.39
aes -> qual	0.14	0.22	0.14	0.16	0.27	0.27	-0.17	1.75	2.59
nav -> qual	0.1	0.12	0.1	0.09	0.28	0.28	-0.21	1.32	3.19
prod -> qual	0.38	0.47	0.38	0.38	0.25	0.25	0.54	0.89	-2.34
pleas -> purch	0.47	0.51	0.47	0.46	0.49	0.49	0.49	0.49	0.74
qual -> purch	0.22	0.18	0.22	0.17	0.50	0.50	0.5	0.5	0.23
System R-squared	n/a	n/a	n/a	n/a	0.22	0.22	n/a	n/a	n/a

surprisingly consistent. John Fox's *sem* package (Fox 2006) yielded smaller estimates than IBM's AMOS package. semPLS and SmartPLS yielded similar estimates. All of the systems of regression equation approaches report regression coefficients rather than correlations; their values tend to be small because all of the data is measured on the same Likert scale. Note that with OLS and WLS where equations are assumed independent, the estimates are in the same range as PLS-PA and covariance methods. When the separate equations share a covariance structure, as they do in SUR, TwoSLS and ThreeSLS, estimators take on a wider range of values, including negative values. The set up of the equations in systems of regression equation approaches make assumptions about causal direction, exogeneity and endogeneity; negative values can imply that these assumptions are violated. No such assumptions are made in SEM where only correlations are computed.

Chapter 5

Data Collection, Control and Sample Size

5.1 The Role of Data

Many questions in social sciences can only be addressed through individual perceptions, impressions and judgments. A consumer's willingness to pay for a product or service is noisy signal, and the consumer has no obligation to follow through on a purchase intent, no matter how much the researcher might like to infer that 'intention' is 'action.' Such inherently unobservable constructs need to be modeled as a latent variables. Personal statements of intent, whether they are for purchases, good deeds or other promises, can only be considered rough indicators; researchers like them because they are cheap and easy to collect by questioning the individual. But like confessions and New Year's resolutions, intentions are pliable and yielding, and often mendacious. Psychologists have created improved polygraph protocols involving such questions over nearly a century, yet polygraph evidence is still not admissible in court. Obtaining truthful and accurate data from surveys and questionnaires is challenging and the quality of information is invariably lacking. Latent constructs that are of actual interest – ones that help us build theory, and which of primary interest – are often unobservable, and the only way to understand them is through objective measurement of related constructs – the indicator variables. Social science data – particularly financial and economic data that are ultimately based on double-entry bookkeeping – is also often highly multicollinear. Different data variables tend not to tell us much that is new about either observed or latent constructs. Double-entry means that by definition, accounting data counts single data times or events multiple times. For example, a single sale will appear as a debit to accounts receivable and credit to sales; later a debit to cash and credit to accounts receivable – one piece of information (the sales event) is turned into four data items. This implies that to get a unambiguous statistical results, we need to sample large numbers of variables, and acquire large datasets of their measurements.

The social sciences are at a disadvantage in data collection when compared with the natural sciences. Historical records like financial statements and surveys of individual behavior tend not only to be subjective, but they are also one-shot, non-applicable measurements. Except in very contrived situations, social scientists find it difficult to set up a controlled laboratory experiment, and rerun it thousands or millions of times. Of all of the social scientists, econometricians are perhaps the most fortunate with respect to datasets. They inherent the masses of data generated by individual, corporate and national accounting – in the US alone, the costs of economic data collection and collation now exceed $1 trillion annually. No other social science comes even close to this level of expenditure on data. Fisher (Fisher, 1935) described statistics as the study of populations, variation and methods of data reduction. Samples need to be reduced to summarize information about the population. The three fundamental tasks of the statistician are to:

1. Define the population
2. Identify the sources of variation
3. Decide how the data should be reduced (simplified as a small number of summary statistics)

The use of randomness in some form allows statisticians to use probability theory – the branch of mathematics that analyses random phenomena; contrasted with statistics that is the discipline of inferring the true state of a population given limited information. One of the fundamental distinctions made in designing statistical studies is whether these will be randomized, or observational studies. In survey sampling one often sees 'convenience samples' (in the accounting profession these are called 'judgmental samples'), implying that random selection procedures have not been used in acquiring the data. The problem is that such samples are not representative of the population (they are only representative of themselves) and the researcher cannot reliably infer anything about the population from the sample. A probability model is required to draw valid inferences from any statistical methods (including SEM methods) – otherwise the conclusions only apply to the data items in the sample.
Observational data may be used to help specify a model (often done in pretest) but is not valid for inferences about the population as a whole. In these situations, precision is less of an issue than bias – the suspicion that researchers have 'cooked the books' to support their prejudices. Despite this, the analysis of non-experimental or quasi-experimental data has captivated statisticians since the fields inception in the 17th century. A useful summary of approaches can be found in (Copas and Li 1997).

5.2 The Ancient Roots of Model-data duality

It is not immediately obvious that data should be distinct and independent from models, no matter how much modern science may be predicated on that assumption. In practice, the human brain's neural networks, programming languages like LISP and APL, Excel spreadsheets, fractals, cellular automata (Wolfram 2002a) and numerous other representations systematically conflate models and data. The control and clarity demanded of scientific argumentation bias research towards a clear model-data dichotomy. This may now be changing with the nascence of computer intensive analytical tools like the methods surveyed for social network analytics in the final chapter of this book. The model-dataset duality is ancient, most famously articulated in Plato's Theory of Forms (also known as the Theory of Ideas, or as Aristotle's hylomorphism). It has been a central tenet of intellectual inquiry throughout history. Plato asserted that abstract (but substantial) ideas, and not the material world of change known through measurement, represented the most fundamental kind of reality. Models are representations of reality that are not directly measurable. Structural equation models reflect this duality explicitly – the structural (inner) model is truly an unmeasurable Platonic form; the measurement (outer) model contains the actual measurements from the material world. In The Republic, Plato (in a dialog with his teacher Socrates) presents 'the allegory of the cave' – a dialog that captures the essence the statistical challenge. In the dialogue, Socrates describes a group of people who have lived chained to the wall of a cave all of their lives, facing a blank wall. The people watch shadows projected on the wall by things passing in front of a fire behind them, and begin to ascribe forms to these shadows. Socrates saw philosophers – the scientists of those days – as 'freed' prisoners who can use their mental tools to perceive reality rather than the shadows. And like the residents of Plato's cave, who know the world only through shadows, researchers cannot completely know the 'true' state of nature – the reality. Rather they have to make do with artificial measurements (shadows lacking dimension and color, and which change shape depending on the angle from which you measure them). Research can only hope to collect a sufficient number of measurements over time to gain some insight into this unseen reality. These measurements are collectively termed data. But they can be unreliable shadows of the things they represent. The hypothetico-deductive model arose in its modern form with the experiments of Galileo Galilei in the 16th century. Growing more complex and reliable over the past five centuries, it has so-far managed to fend off competitors in scientific discourse, and today the hypothetico-deductive model remains perhaps the best-understood theory of scientific method.

The hypothetico-deductive model for scientific inquiry proceeds by formulating a hypothesis in a form that could conceivably be falsified by a test on observable data. This is typically accomplished through the following steps:

1. Specify a model of the real-world based on hunches, experience, exploratory data search, and other subjective means
2. Segment the research model into a series of 'yes/no' questions called hypotheses
3. Conjecture predictions from the hypothesis. Identify the expected characteristics of real-world observations (i.e., the data) you would find were the hypothesis true.
4. Test (i.e., experiment) by collecting evidence (i.e.,data) germane to each hypothesis then applying the tools of inference (i.e., statistics) to draw conclusions.
5. Based on strength, cost and feasibility of additional data collection, revise and improve the experiment until the desired level of certainty about the model truth or falsehood is achieved.

Today, the language of the hypothetico-deductive model is central to the social sciences. But in contrast to the natural science, social science data measurements can be ephemeral and ethereal. Consider for example, a survey of people's intentions to do something – for example a New Year's resolution made on January 1st . Assume that you asked 100 subjects to rank on a scale of 1 (intend to lose zero pounds) to 7 (intend to lose 20 pounds) their intention to lose 20 extra pounds. Now suppose that you used the result of this survey – sample mean and standard deviation of – to predict how many pounds the subjects would lose. You assume that intentions are distributed Normally (even though they are integers that only run from 1 to 7). You predict that the average subject will lose 10 pounds. Do you think that would be an accurate prediction of what subjects actually had done by March 31st? Indeed, given that this is a New Year's resolution, it would not be surprising if a substantial number of subjects had added on weight at the end of three months, or had completely given up any intention of losing weight by that time (the survey measurements make no accommodation for either the intention or reality of adding weight).

Our example suggests that at least two questions are pivotal in dataset choices:

1. Is it representative of the 'population' that is assumed in the model?
2. How much will the collected data cost as a function of quality and quantity?

The first question – "Is it representative of the 'population' that is assumed in the model?" – requires a strategy for data acquisition that is tied to the specific factors or components in the model. Ideally these model factors should be the columns of your dataset – but you usually are not that lucky. In the context of this book, structural equation models give you the luxury of an inner structural model that contains the factors important in the research question, but then allows the collection of data that does not directly measure these structural factors. This ends up being important when models contain many abstract concepts as is common in the social sciences. Still, the data needs to say something about the abstract factors in the model. This is where the representativeness of the dataset becomes important. Model abstractions make sweeping generalizations about, for example, teenagers, consumers, voters, or other such groupings of people. Let's assume that we have a question about teenagers. The total population of teenagers (age 13-19) in the US is about 30 million, simply too large for an affordable dataset. The solution is to sample a representative subset of the population, and extrapolate to the population. But how do we assure that the sample is representative? One approach is to assure that each and every one of those 30 million teenagers has an equal probability of appearing in the sample (much harder than it might first seem, which is the problem the Census Bureau faces every ten years). This is achieved through random sampling strategies. Random sampling avoid selection biases – for example choosing only 19 year old college sophomores because it is easy to find them in classrooms.

In laboratory or agricultural tests that involve testing the effect of treatments, even more involved designs may be concocted to assure representativeness. Experimental design is a separate discipline in statistics that dictates information-gathering exercises where variation is present, whether under the full control of the experimenter or not. The analysis of variation is central to SEM analysis, and the degree to which researchers can control this dictates methodology. Laboratory experiments attempt to cede as much control as possible over treatments and effects to the experimenter. In contrast, natural or quasi-experiments may contain many of the features of laboratory experiments, but are one-shot, non-repeatable experiments. Almost all social science must make do with natural experiments – naturally occurring instances of observable phenomena that approach or duplicate a scientific experiment. In contrast to laboratory experiments, these events aren't created by scientists, but yield scientific data. Natural experiments are a common research tool in fields where artificial experimentation is difficult, such as business, cosmology, epidemiology, and sociology. For example consider the spread of early humans across the Pacific Ocean – and important area

of research for historians. The distribution of populations between islands was essentially random, allowing researchers to treat different groups as independent societies drawn from a common pool. Hypotheses could then be tested in different contexts without fear that an unobserved factor is the cause of differences between island populations The second question – "How much will the collected data cost as a function of quality and quantity?" – requires that we understand the marginal cost of each data item collected. This marginal cost tends to have a substantial impact on cash strapped academics, and dictates experimental design and scope. At the extremes – for example large telescopes, particle accelerators and space flights – data collection may come only once in a lifetime, and a limited quantity available during the research window.
Even without such extreme constraints, there is pressure to economize, or even to cheat, through a variety of shortcuts:

- select data that is convenient (college sophomores) rather than representative (random)
- double and triple count data already collected, for example, through improper use of bootstrapping.
- opportunistically adjust the model to make the dataset seem like it contains more information, through step-wise regression and other techniques to maximize fit statistics without rethinking what the model states about reality
- claim research findings where there exist only compelling patterns, a strategy sometimes seen in data visualization (for example in brain imaging), ordination, simulations and so forth. Methodologies that are useful for data exploration and model specification, for example factor analysis or partial least squares, are not appropriate for testing hypotheses and determining the goodness of fit of data.

These are all essentially ways of gaming the formal methods of science; unfortunately they undermine the credibility of the streams of research in which they are used. They may use the vernacular of statistics to lend credibility to their conclusions, when in fact their conclusions have been chosen in advance, and the statistics are rigged to favor the pre-selected outcomes.

5.3 Data – Model Fit

A major problem that arose in the shift from PCA defined latent variables (formative links proposed by Wold) to researcher defined latent variables (reflective links) is that researchers may get the wrong indicators matched with a particular latent variable – this may generate common factor bias and other biases on the path coefficients. Cronbach's alpha and common factor tests are designed to spot this, but only for a single latent construct. The larger problem is how the information in the indicators aligns with the research model (the latent constructs and paths). There may be information in the dataset on less than the complete set of model latent constructs (e.g., if there are 6 latent constructs, and only 3 PCA components with eigenvalues greater than 1, then there is probably only information in the dataset to support 3 latent constructs – and these may not be exactly the latent constructs chosen by the researcher). The proper way to determine model-data fit is to run a PCA on the data, use the Kaiser criterion to select significant components (those with eigenvalues >1) then see which latent constructs they align with. Then either the model needs to be adjusted to fit the data; or more data (either more observations or more attributes-questions) needs to be collected. Generally if there are two few PCA in the data, more attributes-questions need to be measured. No matter what methodology is used in analysis of the indicator observations, the performance of the researcher's structural model will be explained in terms of its explanatory power relative to some unconstrained optimal partitioning of the data into latent variables. The standard for assessing data model fit for SEM which assume squared error loss, and reveal linear relationships, is commonly set by principal components analysis (PCA) which selects latent variables (components) to maximize the explanation of variance in the data. Latent variables or order – first principal component, second and so forth – in declining proportion of data variance explained. There are actually many ways in which the indicator can be misaligned (in the sense of explaining less variance than a PCA with the same indicators and the same number of latent variables/components). Some of the more insidious problems arise from non-linearity of relationships. These will not show up in SEM, but will confound their results.

As the structural model expands to include more and more latent variables, the opportunities for misspecification increase exponentially. Cronbach's alpha, and other related measures become more and more difficult to interpret, as there will be an exponentially expanded set of ways that indicators can be misassigned to latent variables. This is a weakness in statistics like Cronbach's alpha with can be circumvented by returning to the initial objectives behind the statistic – to resolve the problem in the research's latent constructs, rather than letting a PCA 'organically' select latent variables (principal components) to maximize the variance explained. More fundamentally, it is a problem in the data-model Fit. Researchers set out to answer specific research questions that require definition of a set of concepts – both measurable and abstract. Instruments and studies are designed to collect data, which often comprise the majority of research expenditures. Unfortunately, the data do not always neatly offer up answers to the questions asked of them. Data may be incomplete, or answer different questions than asked, or simply provide insufficient information for an answer. This is not really controllable in advance – it is part of the inherent risk in inquiry. So the researcher can usually be assured that the information in the dataset, and the information required to answer the research questions – in the context of path models, the information required to estimate path coefficients – will not completely coincide. Either the research questions and hypotheses need to be modified to fit the information in the data, or more data needs to be collected. This additional data collection can acquire more indicators (information on latent variables that is missing in the original dataset) or can acquire more observations (allowing more precise estimates, or the identification of smaller effects).

In SEM, the connections between latent variables and indicator variables is referred to as measurement or outer model. A model with all arrows pointing outwards is called a reflective measurement model – all latent variables have reflective measurements. A model with all arrows pointing inwards is called a formative measurement model – all latent variables have formative measurements. A model containing both, formative and reflective latent variables is referred to as multi-block model. The direction of arrows are commonly used to distinguish reflective from formative links, but this is sometimes misinterpreted as distinguishing causal direction. In fact, any assertion of causal direction in indicator-latent variable links would be dubious, since the indicators are observed and the latent variables are unobserved. These two approaches differ philosophically by their assumptions about exogeneity. In formative measurement, the latent variable is presumed to be a linear combination of exogenous indicator variables. In reflective measurement, the indicator variables reflect the values of an unobservable exogenous latent variable which we assume can be represented as a linear function of the indicator variables. Mathematically, the exogenous variable is placed on the right-hand side of the model equations, so weights and estimators may vary depending on whether you decide latent or indicator variables are exogenous.

In practice, we may consider reflective links to be model driven – they are created in the design of the research study. In the case of reflective indicator links, typical of classical factor analysis models, the latent constructs give rise to observed variables that co-vary among them and the model aims at accounting for observed variances or covariances. This is typically what is encountered in surveys, where clusters of questions are intended to glean information about a particular unobservable (latent) construct. The researcher creates a cluster of reflective variable links around a latent variable in the construction of the model and the survey instrument. Formative links are data driven – they are inferred *ex posteriori* from data that is already collected. Formative indicator links are emergent constructs – combinations of observed indicators and are not designed to account for observed variables. The components of a typical PCA analysis are such combinations of observed indicators – these components don't necessarily correspond to specific constructs, observed or not. The researcher creates a cluster of formative variable links around a latent variable through factor analysis or principal component analysis computed with a mathematical objective in mind, e.g., minimizing variance, dimensional reduction, and so forth.

Ideally we would like for formative and reflective indicator links to be identical. A successful experimental design will preselect indicator variables for each latent construct that are strongly correlated with each other – which are multicolinear. This helps validate the assertion that each of several questions on a survey, or measurements in an experiment, represent exactly the same thing – measurements in an unobserved latent construct. In implementation, a latent variable is simply a linear function of the indicators and the latent variable exists in concept only as the realized value of this linear function. This is the reasoning behind the three tools for insuring that the indicator variables are informative, consistent and appropriate surrogates for the unobserved latent variable:

1. Harman's single factor test and Cronbach's alpha are concepts in the validation of reflective links;
2. The related concept of the Kaiser criterion is used in the choice of indicators in formative links.

A veritable cottage industry has evolved to generate related vernacular; fortunately, much of this is unnecessary as the basic concepts are straightforward and simple. Common factor bias is the same as common method variance – 'factor' means the same thing as 'indicator,' and 'method' is the way in which this indicator is chosen. Bias and variance are used similarly as well, to indicate the implied changes in the latent variable. Technically, these are portions of the model variation that can be ascribed to problems in the measurement method (e.g., the way in which the survey instrument was constructed; whether the questions were 'on scale' and 'balanced') rather than the (latent) constructs of interest. These are presumed to

Table 5.1: Guidelines for Assessing Internal Consistency Value of Cronbach's alpha

Range.of.alpha	Internal.consistency
.9 to 1.0	Excellent
.8 to .9	Good
.7 to .8	Acceptable
.6 to .7	Questionable
.5 to .6	Poor
Below .5	Unacceptable

cause systematic measurement error and bias the estimates of the 'true' relationship among theoretical constructs. In biasing the model path estimates, these can lead to problems in hypothesis tests.

Where confirmation is the objective, one problem that can arise in building the structural model completely without reference to the data (i.e. all reflective indicator links) is that the latent constructs chosen by the researcher may be substantially different than those that would drop out of an exploratory factor analysis. Harman's one factor test (Podsakoff and Organ 1986) has been suggested as a test for common factor bias or common method variance. The most common reason that this test is needed is that the model is constructed without reference to clustering in the underlying data; i.e., it is entirely theory-driven (not in itself a bad thing).

The Harman single-factor test performs a factor analysis or principal component analysis on all of the indicators collected for the study (presumably these will all be reflective indicator links when you use this test) and assesses whether the very first factor or component is significantly larger than all of the other components. If so, it is assumed that your unobserved latent variables are multicolinear, and there is a 'common method bias' – i.e., a single unobserved factor, perhaps introduced in the survey or experimental design indicative of common method variance – that influences all of the latent variables and overstates the correlations between them.

Conversely, one can inspect each latent variable and its associated cluster of indicators to see if these are appropriate choices for reflective indicators. Cronbach's alpha is used as a coefficient of reliability for such choices – it provides a measure of the internal consistency or reliability of a psychometric test score on questionnaires. It was first named (alpha) by Lee Cronbach in 1951, as he had intended to continue with further coefficients. Cronbach's alpha statistic is widely used in the social sciences, business, nursing, and other disciplines. It attempts to answer the extent to which questions, test subjects, indicators, etc measure the same thing (i.e. latent construct). (Nunnally, Bernstein, and others 1967) provides the guidelines for assessing internal consistency using Cronbach's alpha

5.4 Latent Variables

Latent variables are in practice typically constructed from linear combinations of indicator variables. We would like each latent variable to represent a unique and meaningful construct, even though it is unobserved. And we would like to collect sufficient indicator data on each of the latent constructs that we include in the model – put differently, we would like the data items we have collected to coalesce, under some clustering algorithm, to clusters that match with the latent variables. There are two ways to do this: 1. We can collect the data, run a principal component analysis (or other factor analysis) on that data, and include latent constructs for the most significant components derived from the principal component analysis. The Kaiser criterion would suggest that all components with eigenvalues over one be included. Indicator variables and links are formative in this case. 2. We can build a theory based model, and then collect data to test the model (and thus indirectly, the theory). This requires two additional steps over the first approach: a. The experiment or survey instrument need careful construction around the model parameters. Data scaling, granularity, location and reliability are all confounding issues in correctly constructing survey instruments. b. Once the data is collected, it needs to be tested to make sure that the clusters that actually exist in the data correspond to the expected clustering (i.e., the reflective indicator-link constructs for each predefined latent construct). Harman's one factor test and Cronbach's alpha are generic tests In the model driven case where reflective indicator-link constructs are built from theory, the researcher takes on an additional obligation – to assure that data is collected for each latent construct, and that this data is reliable, consistent and adequate to support the conclusions of the research. Cliff (Cliff, 1988) argues that using both Cronbach's alpha and the Kaiser criterion to identify components with significant eigenvalues is required to properly validate the adequacy and reliability of the data. These two related tests need to be used together to assess validity of indicator-latent variable clustering. Though (Kaiser 1960) argues for slightly differing criteria, it is clear that this expanded notion of principal components testing for data-model fit was on the minds of both Kaiser and Cronbach when they developed their assessments in the 1950s. Problems of data-model fit – whether you are discussing common factor bias, interfactor reliability, or some other criterion – can be avoided a priori through a pretest of the clustering of indicator data. Common factor bias occurs because procedures that should be a standard part of model specification are in practice left until after the data collection and confirmatory analysis. Jöreskog developed PRELIS for these sorts of pretests and model re-specifications. If this clustering shows that the indicators are providing information on fewer variables than the researchers' latent SEM contains, this is an indication that more indicators need to be collected that will provide:

1. additional information about the latent constructs that don't show up in the cluster analysis; and
2. additional information to split one exploratory factor into the two or more latent constructs the research needs to complete the hypothesized model.

In exploratory factor analysis, the two tests that are most useful for this are the Kaiser criterion that retains factors with eigenvalues greater than one (unless a factor extracts at least as much information as the equivalent of one original variable, we drop it) and the scree test proposed in (Cattell 1966) that compares the difference between two successive eigenvalues and stops taking factors when this drops below a certain level. In either case, the suggested factors are not necessarily the latent factors that the researcher's theory would suggest – rather they are the information that is actually provided in the data, this information being the main justification for the cost of data collection. So in practice, either test would set a maximum number of latent factors in the SEM if that SEM is to be explored with one's own particular dataset.

5.5 Linear Models

All four of the methods presented in this book – correlation, PLS path analysis, covariance structure methods and systems of equations regression – use linear models and generalizations. These are appropriate where the population characteristics are linear; but they are misleading where they are not. Many real-world relationships are non-linear – not just a little, but substantially non-linear. For example, the technology acceleration depicted by Moore's Law (computing power doubles every 18 months) is exponential; the value of social networks as a power of the number of members; the output of a factory has declining returns to scale. It is important to always look for a physical model underlying the data. Assume a linear model as a starting point only or a simplification which may be useful, but which cannot go unexamined. Models may be conceived and used at three levels. The first is a model that fits the data. A test of goodness-of-fit operates at this level. Linear models often fit the data (within a limited range) but do not explain the data. A second level of usefulness is that the model predicts future observations – it is a forecast model. Such a model is often required in screening studies or studies predicting outcomes. A third level is that a model reveals unexpected features of the situation being described – it is a structural model.

5.6 Hypothesis Tests and Data

Classical (Neyman-Pearson) hypothesis testing requires making assumptions about underlying distributions from which the data were sampled. Usually the only difference in these assumptions between parametric and nonparametric tests is the assumption of a Normal distribution; all other assumptions for the parametric test apply to the nonparametric counterpart as well. In SEM, while PLS path analysis does not make distributional assumptions, the covariance and simultaneous equation regression approaches both assume Normal datasets; nonparametric SEM approaches do not exist. Hypothesis testing very often assumes that data observations are:

1. Independent
2. Identically distributed (come from the same population with the same variance)
3. Follow a Normal distribution

These assumptions are made both for convenience and for tractability; and for simple models they may be good enough. But at a minimum, the researcher is obligated, prior to model fitting, to test the dataset to assure that data are independent, and that they represent the population. This is often accomplished through various exploratory tests, such as histograms of observations. The third assumption tends to be a substantial hurdle in survey research where responses are recorded on a Likert scale, which is discrete, truncated and categorical, where a Normal distribution offers a poor approximation.

5.7 Data Adequacy in SEM

5.7.1 Does our dataset contain sufficient information for model analysis?

The complex, networked structures of SEM create significant challenges for the determination of sample size and adequacy. From a practical viewpoint, sample size questions can take three forms:

1. *A priori*: will ask what sample size will be sufficient given the researchers prior beliefs on what the minimum effect is that the tests will need to detect.
2. *Ex posteriori*: will ask what sample size should have been taken in order to detect the minimum effect that the researcher actually detected in an existing (either sufficient or insufficient) test. If the *ex posteriori* measured effect is smaller than the researchers prior beliefs about the minimum effect then sample size needs to be increased commensurately.
3. Sequential test optimal-stopping: is couched in a sequential test optimal-stopping context, where the sample size is incremented until it is considered sufficient to stop testing.

In addition, not all sample points are created equal. A single sample data point copied three times over still has only one single sample data point worth of information. Even where complex 'bootstrapping' processes are invoked to duplicate sample points, it is doubtful whether new information about a population is actually created (and where it is, it might better be injected into the data through Bayesian methods, or aggregation). If our research question is about the wealth of a consumer group, then a dataset of colors of the sky at different times of the day will not provide information relevant to the research question. Sample data points will contain differing amounts of information germane to any particular research question. Several datapoints may contain information that overlaps, which is one cause of multicolinearity. The distribution of random data may also differ from modeling assumptions, a problem that commonly occurs in the SEM analysis of Likert scale survey data. To this day, methodologies for assessing suitable sample size requirements remain a vexing question in SEM based studies. The number of degrees of freedom consuming information in structural model estimation increases with the number of potential combinations of latent variables; while the information supplied in estimating increases with the number of measured parameters (i.e., indicators) times the number of observations (i.e., the sample size) – both are non-linear in model parameters. This should imply that requisite sample size is not a linear function solely of indicator count, even though such heuristics are widely invoked in justifying SEM sample size. Monte Carlo simulation in this field has lent support to the non-linearity of sample size requirements, though research to date has not yielded a sample size formula suitable for SEM.

Structural equation modeling in the social sciences has taken a casual attitude towards choice of sample size. Since the early 1990s, social science researchers have alluded to an ad hoc rule of thumb requiring the choosing of 10 observations per indicator in setting a lower bound for the adequacy of sample sizes. Justifications for this "rule of 10"" appear frequently without reference to the original articulation of the rule in(Nunnally, Bernstein, and others 1967) where it was suggested (without providing supporting evidence) that in estimation 'a good rule is to have at least ten times as many subjects as variables.' (D. L. Goodhue, Lewis, and Thompson 2012a), (D. L. Goodhue, Lewis, and Thompson 2012b), (Goodhue, Lewis, and Thompson 2006) studied the rule of 10 using Monte Carlo simulation finding that PLS-PA analysis had inadequate power to detect small or medium effects at small sample (see also (Goodhue 1995) ans (Goodhue and Thompson 1995)). This finding was not unexpected, as similar PLS-PA rules of thumb have been investigated since (Nunnally, Bernstein, and others 1967). (K. A. Bollen 1989) stated that "though I know of no hard and fast rule, a useful suggestion is to have at least several cases per free parameter" and (P. M. Bentler and Mooijaart 1989) suggested a 5:1 ratio of sample size to number of free parameters. But was this the right question? Typically their parameters were considered to be indicator variables in the model, but unlike the early path analysis, structural equation models today are typically estimated in their entirety, and the number of unique entries in the covariance matrix is $\frac{p(p+1)}{2}$ when p is the number of indicators. It would be reasonable to assume that the sample size is proportional to $\frac{p(p+1)}{2}$ rather than p. Unfortunately, Monte Carlo studies conducted in the 1980s and 1990s showed that the problem is somewhat more subtle and complex than that, and sample size and estimator performance are generally uncorrelated with either $\frac{p(p+1)}{2}$ or p.

Difficulties arise because the p indicator variables are used to estimate the k latent variables (the unobserved variables of interest) in SEM, and even though there may be $\frac{p(p+1)}{2}$ free parameters, these are not individually the focus of SEM estimation. Rather, free parameters are clustered around a much smaller set of latent variables that are the focus of the estimation (or alternatively, the correlations between these unobserved latent variables are the focus of estimation). (Tanaka 1987) argued that sample size should be dependent on the number of estimated parameters (the latent variables and their correlations) rather than on the total number of indicators; a view mirrored in other discussions of minimum sample sizes (M. W. Browne and Cudeck 1989). . (H. W. Marsh and Bailey 1991) concluded that the ratio of indicators to latent variables rather than just the number of indicators, as suggested by the rule of 10, is a substantially better basis on which to calculate sample size, reiterating conclusions reached by (Boomsma 1982) who suggested using a ratio $r = \frac{p}{k}$ of indicators to latent variables. Information input to the SEM estimation increases both with more indicators per latent variable, as well as with more sample observations. A series of studies (Ding, et al. 1995) found that the probability of rejecting true models at a significance level of 5% was close to 5% for $r = 2$ (where r is the ratio of indicators to latent variables) but rose steadily as r increased – for $r = 6$, rejection rates were 39% for sample size of 50; 22% for sample size of 100; 12% for sample size of 200; and 6% for sample size of 400. (Boomsma 1982)'s

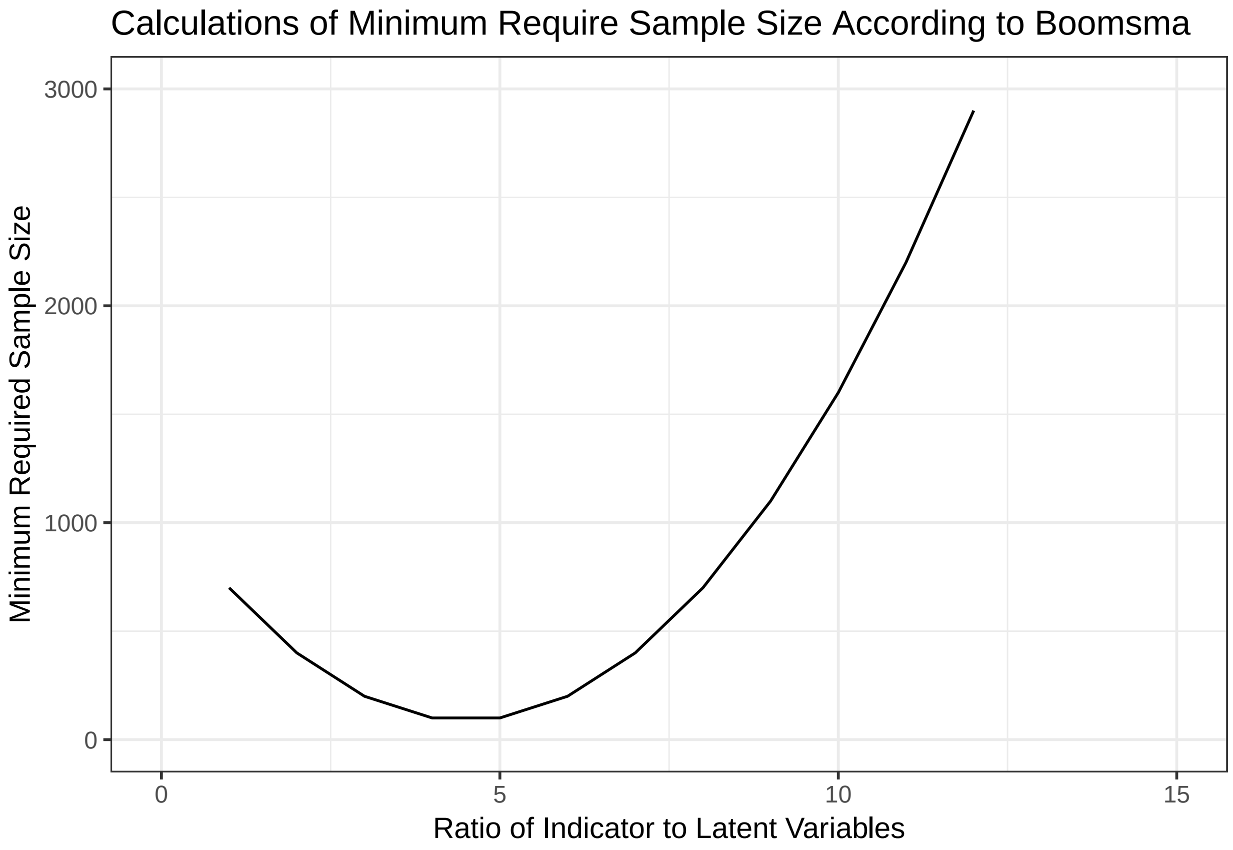

Figure 5.1: Calculations of Minimum Require Sample Size According to Boomsma

simulations suggested that a ratio of indicators to latent variables of $r = 4$ would require a sample size of at least 100 for adequate analysis; and for $r = 2$ would require a sample size of at least 400. (H. W. Marsh and Bailey 1991) ran 35,000 Monte Carlo simulations on LISREL CFA analysis, yielding data that suggested that: would require a sample size of at least 200; would require a sample size of at least 400; $r = 12$ would require a sample size of at least 50. Consolidation and summarization of these results suggest sample sizes: $n \geq 50r^2 - 450r + 1100$ where r is the ratio of indicators to latent variables. Furthermore, (H. W. Marsh and Bailey 1991) recommends 6 to 10 indicators per latent variable, assuming 25-50% of the initial choices add no explanatory power, which they found to often be the case in their studies. They note that this is a substantially larger ratio than found in most SEM studies, which tend to limit themselves to 3-4 indicators per latent variable. It is possible that a sample size “rule of 10”" observations per indicator may indeed bias researchers towards selecting smaller numbers of indicators per latent variable in order to control the cost of a study or the length of a survey instrument. The following section depicts the sample size implied in Boomsma’s simulations.

Boomsma’s guideline is an improvement on the *rule of 10*, but it is not sufficient for hypothesis testing, because it fails to take into account significance and power of the test, minimum detectable effect, and scaling.

5.7.2 The “Black-box” problem

Latent structural models historically suffered from some of the same problems that machine learning models are currently facing. They produce solutions to data analysis questions involving unobserved phenomena, but seem to lack clear guidelines on the sufficiency of data required to support any given research conclusion − SEM is a “black-box” that provides useful analyses, but with poorly defined rules for usage. This has been problematic, because funding agencies, ethics boards and research review panels frequently request that a researcher perform a power analysis. Their argument is that if a study is inadequately powered, there is no point in completing the research. Additionally, in the framework of SEM the assessment of power is affected by the ambiguity of information contained in social science data.

Westland (2010) developed an algorithm for computing the lower bound on sample size in SEM, the best that researchers could rely on imprecise simulations or unsubstantiated “rules of thumb” to gauge sample size adequacy. Traditional Neyman-Pearson hypothesis testing methods required to confirm or reject the existence of a minimum effect in statistical tests given significance and power levels were simply not available for models with SEM’s complexity. In this section we review the sample size computation presented in (Westland 2010) which answers the research question “What is the lower bound on sample size for confirmatory testing of SEM as a function of these minimum effect size, significance and power of the test?” We want to detect a minimum correlation δ (effect) in estimating k latent (unobserved) variables, at given significance and power levels $(\alpha^*, 1-\beta)$. Our solution will be a function in the form $n = f[k, \delta \mid \alpha^*, \beta]$. We initially adopt the standard targets for our required Type I and II errors under Neyman-Pearson hypothesis testing of $\alpha^* = 0.05$ and $\beta = .20$; but these requirements can be relaxed for a more general solution. Structural equation models are characterized here as a collection of pairs of canonically correlated latent variables, and adhere to the standard normalcy assumption on indicator variables. This leads naturally to a deconstruction of the SEM into an overlapping set of bivariate normal distributions. It is typical in the literature to predicate an SEM analysis with the caveat that one needs to make strong arguments for the complex models constructed from the unobserved, latent constructs tested with the particular SEM, in order to support the particular links that are included in the model. This is usually interpreted to mean that each proposed (and tested) link in the SEM needs to be supported with references to prior research, anecdotal evidence and so forth. This may simply mean the wholesale import of a preexisting model (e.g., ‘theory or reasoned action’ model or ‘technology acceptance model’) based on the success of that model in other contexts, but not specifically building on the particular effects under investigation. But it is uncommon to see any discussion of the particular links (causal or otherwise) or combinations of links that are excluded (either implicitly or explicitly) from the SEM model. Ideally, there should also be similarly strong arguments made for the inapplicability of omitted links or omitted combinations of links.

We can formalize these observations by letting i be the number of the potential links (canonical correlations) between latent variables. Extend the individual link minimum sample size to a minimum sample size for the entire SEM; building up from pairs of latent variables by determining the number of possible combinations of the i pairs, each with an ‘effect’ that needs detection. The problem which (Westland 2010) solved using combinatorial analysis, is to compute the number of distinct structural equation models that can exist in terms of the set of link-correlations that are more than the stated minimum effect size. Our desired sample size is large enough to just reduce our number of choices to one − the hypothesized model.

Assume that a link on which the effect is hypothesized to be significant is a 1, and on a link where the effect is not significant, the link is a 0. Then each combination $[0, 1]$ links requires us to discriminate among a set of $\frac{k(k-1)}{2}$ unique combinations of latent variables. The unique model hypothesized in any particular study will be one of $2^{\frac{k(k-1)}{2}}$ ways of connecting these latent variables; testing must discriminate this path from the possible $\frac{k(k-1)}{2} - 1$ other paths which collectively define the alternative hypothesis.

For hypothesis testing with a significance of α^* on each link, it is necessary to correct for effective significance level α in differentiating one possible model from all other hypothesized structural equation models that are possible. The Šidàk correction is a commonly used alternative for the Bonferroni correction where an experimenter is testing a set ofhypotheses with a dataset controlling the family-wise error rate. In the context of the current research the Šidàk correction provides the most accurate results. For the following analysis,a Sidak correction gives $\alpha = \alpha(k) = 1 - (1 - \alpha^*)^{\frac{2}{k(k-1)}}$.

5.7.3 Minimum effect size and correlation metrics

Minimum effect, in the context of structural equation models, is the smallest correlation between latent variables that we wish to be able to detect with our sample and model. Small effects are more difficult to detect than large effects as they require more information to be collected. Information may be added to the analysis by collecting more sample observations, by adding parameters, and by constructing a better model. Sample size for hypothesis testing is typically determined from a critical value that defines the boundary between the rejection (set by) and non rejection (set by) regions. The minimum sample size that can differentiate between and occurs where the critical value that is exactly the same under the null and alternative hypotheses. The approach to computing sample size here is analogous to standard univariate calculations but using a formulation for variance customized to this problem. In the context of structural equation models, canonical correlation between latent variables should be seen simply as correlation, the ‘canonical’ qualifier referring to the particulars of its calculation in SEM since the latent variables are unobserved, and thus cannot be directly measured. Correlation is interpreted as the strength of statistical relationship between two random variables obeying a joint probability distribution (Kendall and Gibbons 1990) like a bivariate normal. Several methods exist to compute correlation: the Pearson’s product moment correlation coefficient (Fisher 1921) Spearman’s rho and Kendall’s tau (Kendall and Gibbons 1990) are perhaps the most widely used (Samuel, Mari, and Kotz 2001). Besides these three classical correlation coefficients, various estimators based on M-estimation (Shevlyakov and Vilchevski 2002) and order statistics (Schechtman and Yitzhaki 1999) have been proposed in the literature. Strengths and weaknesses of various correlation coefficients must be considered in decision making. The Pearson coefficient, which utilizes all the information contained in the variates, is optimal when measuring the correlation between bivariate normal variables

(Stuart, Kendall, and Ord 1991). However, it can perform poorly when the data is attenuated by nonlinear transformations. The two rank correlation coefficients, Spearman's rho and Kendall's tau, are not as efficient as the Pearson correlation under the bivariate normal model; nevertheless they are invariant under increasing monotone transformations, thus often considered as robust alternatives to the Pearson coefficient when the data deviates from bivariate normal model. Despite their robustness and stability in non-normal cases, the M-estimator-based correlation coefficients suffer great losses (up to 63% according to (Xu et al. 2010)) of asymptotic relative efficiency to the Pearson coefficient for normal samples, though such heavy loss of efficiency might not be compensated by their robustness in practice. (Schechtman and Yitzhaki 1999) proposed a correlation coefficient based on order statistics for the bivariate distribution which they call Gini correlation (because it is related to Gini's mean difference in a way that is similar to the relationship between Pearson correlation coefficient and the variance). As a measure of such strength, correlation should be large and positive if there is a high probability that large or small values of one variable occur (respectively) in conjunction with large of small values of another; and it should be large and negative if the direction is reversed. We will use a standard definition of minimum effect size to be detected – the strength of the relationship between two variables in a statistical population as measured by the correlation for paired latent variables – following conventions articulated in (Nakagawa and Cuthill 2007). Where we are assessing completed research, we can substitute for δ the smallest correlation (effect size) on all of the links between latent variables in the SEM.
(Xu et al. 2010) used Monte Carlo simulations to verify the accuracy and robustness of Gini correlations in bivariate normal distribution estimation, using asymptotic relative efficiency and root mean square error performance metrics) showing that they are applicable for data of even relatively small sample sizes (down to around 30 sample points). Their simulations confirmed and extend (He and Nagaraja 2011) Monte Carlo simulations exploring the behavior of nine distinct correlation estimators of the bivariate normal correlation coefficient. The Gini estimator was found generally to reduce bias and improve efficiency as well or better than other correlation estimators in the study. (Xu et al. 2010) also compared with three other closely related correlation coefficients:

- classical Pearson's product moment correlation coefficient,
- Spearman's rho, and
- order statistics correlation coefficients.

Gini correlation bridges the gap between the order statistics correlation coefficient and Spearman's rho, and its estimators are more mathematically tractable than Spearman's rho, whose variance involves complex elliptic integrals that cannot be expressed in elementary functions. Their efficiency analysis showed that estimator 's loss of efficiency is between 4.5% to 11.3%, much less than that of Spearman's rho which ranges from 8.8% to 30.5%.

5.7.4 Minimum Sample Size for SEM Model Analysis

(Westland 2010) provided a general formula for computing sample size required to test a given SEM model to explain a particular dataset. (Westland 2010) provided *necessary* but not sufficient conditions for sample size – as such the formula sets an absolute lower bound on sample size for any given model-data pair. Figures @ref(fig:westland_lv), @ref(fig:westland_effect), @ref(fig:westland_power) and @ref(fig:westland_sig) show the effect on minimum required sample size $n = f[k, \delta \mid \alpha^*, \beta]$ of the test choice parameters $k, \delta, \alpha^*, \beta$ – the *number of latent variables*, *minimum effect size*, *significance level at each link*, and *power of test* respectively.

```
## R language code to compute and plot SEM sample size

library(ggplot2)
westland <- data.frame()  ## data.frame for ggplot
for (i in 1:10)  {
lv <- 2*i           ## number of latent variables (k)
effect <-.1        ## minimum acceptable effect size (delta)
power <- .8        ## power of test
sig_star <- .05    ## significance of test at each link
sig <-  1 - (1 - sig_star)^(2/(lv*(lv-1)))   ## Sidak correction
rho <- effect    ## links' correlation > minimum effect
A <- 1 - rho^2
B <- rho*asin(rho/2)
C <- rho*asin(rho)
D <- A/((3-A)^.5)
H <- (effect/(qnorm(1-sig/2)-qnorm(1-power)))^2
n <- ceiling(
  1/(2*H)*
    (A*(pi/6-B+D)+H+
       ((A*(pi/6-B+D)+H)^2+
          4*A*H*(pi/6+A^.5+2*B-C-2*D)
        )^.5) )

  westland[i,1] <- lv
  westland[i,2] <- n
}
ggplot(data = westland, aes(x = westland[,1], y = westland[,2])) +
  stat_smooth(se=F,formula=y ~ log(x))+
  ggtitle("effect =.1,  power = .8, significance =  .05")+
  xlab("Number of Latent Variables")+
  ylab("Minimum Required Sample Size")+
  ylim(800,2500) +  xlim(0,20)
```

The formula in (Westland 2010) is not specific to any particular approach (PLS-PA, covariance, or regression) but is a consequence of the link structure of the SEM network. (Westland 2010)'s sample size is not sufficient to assure sample adequacy because so many other factors can affect estimation – multicolinearity, appropriateness of data sets, and so forth. Additionally, the information contained in the sample and indicator variables must be adequate to compensate for variations in particular SEM estimation methodologies. (Jöreskog and Goldberger 1975)

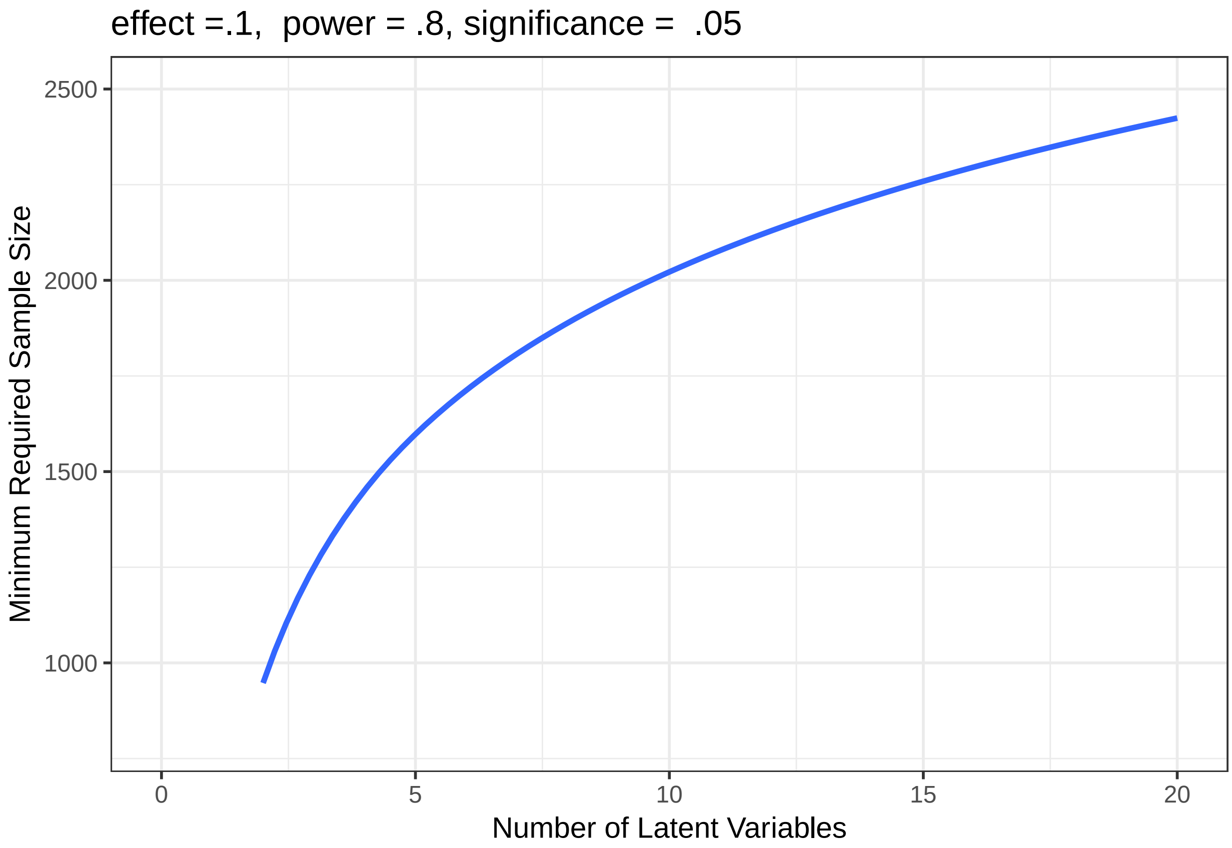

Figure 5.2: (#fig:westland_lv)Number of Latent Variables and Minimum Required Sample Size

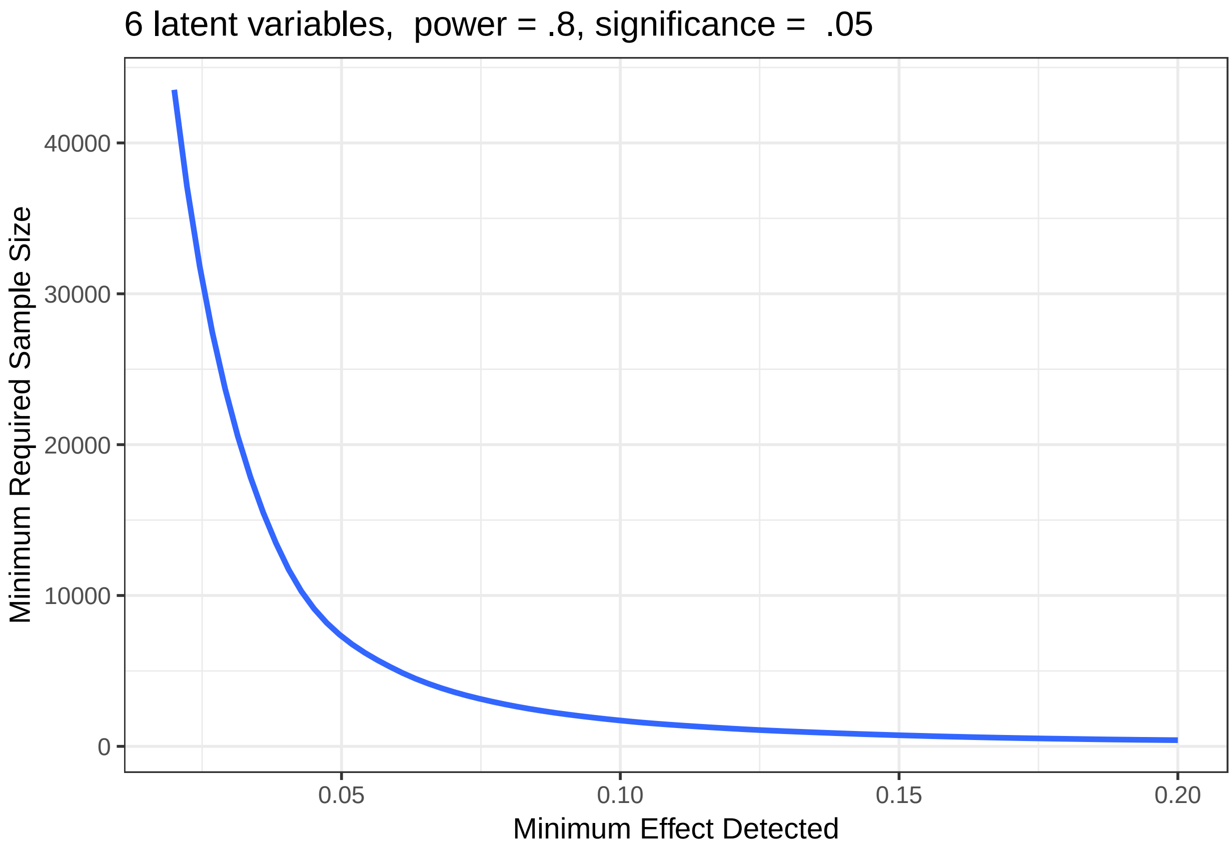

Figure 5.3: (#fig:westland_effect)Effect Size and Minimum Required Sample Size

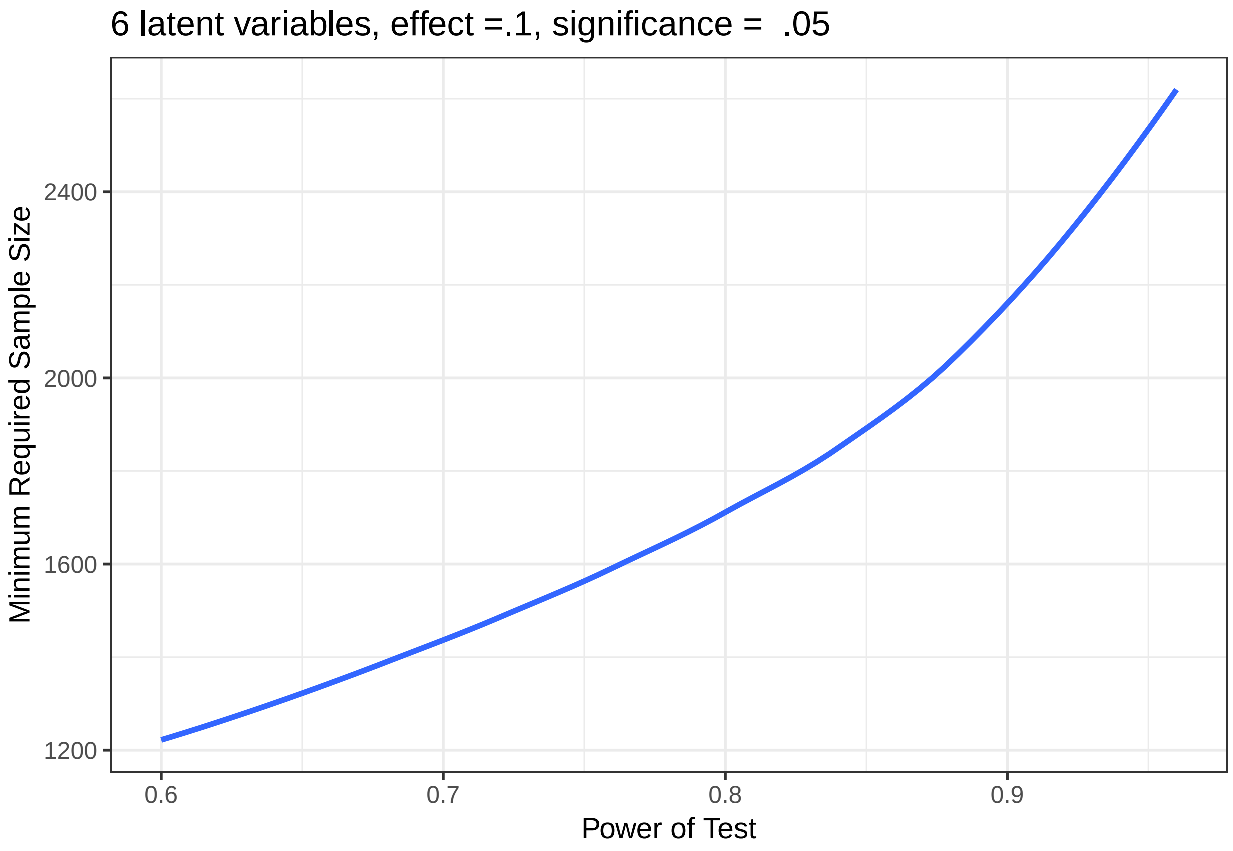

Figure 5.4: (#fig:westland_power)Power of Test and Minimum Require Sample Size

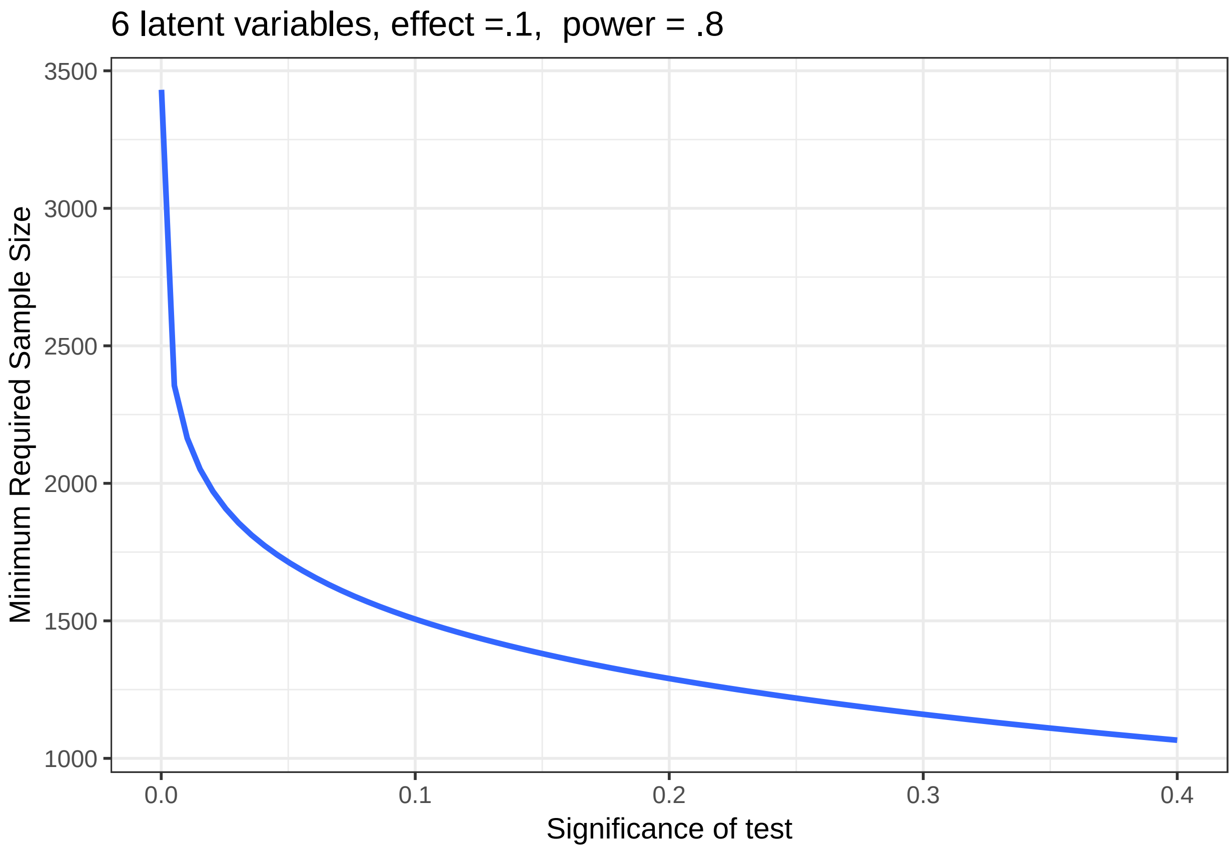

Figure 5.5: (#fig:westland_sig)Significance of Test and Minimum Require Sample Size

suggests that non-normal indicators require one, two or three orders of magnitude magnitudes larger samples, depending on distribution. (Dhrymes 1974) demonstrate that the IV/2SLS techniques converge to the same estimators as PLS-PA and covariance methods, but are more robust than PLS-PA and thus may demand smaller samples. In addition, Likert-scaled data is clearly not normal, since it is truncated at both the upper and lower ranges, and is discrete.

5.8 Can Resampling recover information lost through Likert mapping?

The loss of information in the Likert mapping from a continuous set of beliefs into a very simple, discrete Likert categorical distribution is likely to be both substantial, as well as difficult to measure. Subjects may not have strongly held beliefs, and where they do, these beliefs can change quickly under the influence of new data. Resampling or "bootstrapping" is one method – but not one without controversy – for attempting to gather more information about the actual perceptions being approximated in the Likert scale.

The statistical use of 'bootstrapping' was borrowed from *The Surprising Adventures of Baron Munchausen* (Raspe 2009) where Baron Munchausen pulls himself out of a swamp by his bootstraps. Both bootstrapping and jackknifing serve an important role is assuring the stability of estimators, or determining where more data may need to be collected. Bootstrapping algorithms are built into most SEM software, making them relatively easy to misapply when drawing conclusions from a dataset. They have been used incorrectly to assert the desirable small sample properties of SEM, and in particular to force an inadequate dataset to support an incorrect hypothesis Resampled data provides a simple, straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution, such as are typical of SEM. Like Monte Carlo approaches, it is an appropriate way to control and check the stability of the results, but is dependent on distributional assumptions that may or may not be true. When applied properly, it is a useful tool for developing and incorporating model assumptions that are consistent with the data. Bootstrap data is asymptotically consistent, but does not provide general finite-sample assurances. Moreover, there is no guarantee that bootstrapped Likert data will better represent underlying beliefs than the original data. Bootstrapping does not create new information; if the researcher is lucky, it may provide modeling insights that were not previously available – somewhat like a pretest. The basic idea behind bootstrapping is that the sample we have collected is often the best guess we have as to the shape of the population from which the sample was taken. Thus we could assume (without actually collecting data) that future data will come from this empirical distribution, and artificially generate more data. We can use similar ideas for imputation of missing data, and all of this falls under the broader rubric of resampling. Rather than making assumptions directly – for example, that the data is drawn from a Normal population – we can let the bootstrap generated data introduce this assumption into the modeling. It estimates the sampling distribution of an estimator by sampling with replacement from the original sample, most often with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter like a mean, median, proportion, and so forth. Jackknifing is similar to bootstrapping, and is used in statistical inference to estimate the bias and standard error of a statistic, when a random sample of observations is used to calculate it. The basic idea behind the jackknife variance estimator lies in systematically recomputing the statistic estimate leaving out one or more observations at a time from the sample set.

5.9 Data screening

Prior to the descriptive and positive testing, it is important to screen the data. This is the first step towards formalization of the research. The order of the screening is important as decisions at the earlier steps influence decisions to be taken at later steps. For example, if the data is both non-normal and has outliers, the decision to delete values or transform the data is confronted. Transformation of the variable is usually preferred as it typically reduces the number of outliers, and is more likely to produce normality, linearity, and homoscedasticity; in social science work, data is often limited to positive values only, and may be ordinal as well, as is the case for Likert scale responses. Screening will aid in the isolation of data peculiarities and allow the data to be adjusted in advance of further multivariate analysis. (Tabachnick and Fidell 2007) suggest the following data screening tasks:

1. Inspect univariate descriptive statistics for accuracy of input
    a. out-of-range values, be aware of measurement scales
    b. plausible means and standard deviations
    c. coefficient of variation
2. Evaluate amount and distribution of missing data: deal with problem
3. Independence of variables
4. Identify and deal with non-Normal variables
    a. check skewness and kurtosis, probability plots
    b. transform variables (if desirable)
    c. check results of transformations
5. Identify and deal with outliers
    a. univariate outliers
    b. multivariate outliers
6. Check pairwise plots for nonlinearity and heteroscedasticity
7. Evaluate variables for multicolinearity and singularity
8. Check for spatial autocorrelation

Fortunately, there are excellent tools available to the researcher for data screening. PRELIS (a part of the LISREL statistical package but which can be made to work with PLS packages) and its SPSS counterparts (which are often used by themselves or for preliminary testing, transformation and culling of data in PLS) in AMOS – these provide automated support for all of these screening tasks. The R-language statistical package supports many data screening tasks, that allows for ‘eyeballing’ data for patterns, and transforming, plotting residuals, and other useful functions on the fly. Without proper data screening, data exploration and testing later in the research cycle cannot be relied upon to provide reliable answers to research questions.

5.10 Exploratory Specification Search

(Leamer 1978) surveyed statistical specification search approaches (in his case, Bayesian methods) that can be used for preliminary model building, model re-specification and ad hoc inference with weak and / or non-experimental data. Such specification searches are the main task of hypothesis justification, pretests and suggestions for extending already completed research. They play an especially important role in the social sciences where core constructs are often not directly observable (e.g., consumer satisfaction) and were two important classes of data – survey and economic – are non-experimental. Iteratively combining:

1. factor analysis and other cluster analysis methods, with,
2. PLS-PA for holistically exploring the causal relationships between 'clusters' (the first step towards generating theory-driven latent factor constructs), and
3. step-wise regression for decisions on whether or not to include factors, provides comprehensive exploration of the parameter space supported by a particular set of data and prior beliefs.

Clearly the results cannot be seen as ends in themselves. Rather this exploratory specification search assures – in moving to the rigorous confirmatory testing of a particular model – that any feasible models, variables and causal relationships that should be tested are included. When SEM are built around valid real world constructs (even if these are unobservable) the algorithms proposed in this paper impose only weak additional assumptions on the indicators and latent variables in order to compute sample sizes adequate for estimation. Our limited application to a window of IS and e-commerce publications has shown that concerns are warranted concerning existing SEM sample size calculations and we need to remain suspicious of conclusions reached in studies based on inadequate sample sizes. Furthermore, a large number of studies in our sample devised their tests without first committing to minimum effect size that they were trying to detect, or indicated in portion of non-response in surveys. It is clear that journal referees need to begin asking for survey response, minimum effect size and a justification of the sample size. By incorporating these suggestions, it is argued that the research community will enhance the credibility and applicability of their research, with a commensurate improved impact and influence in both industry and academe.

Chapter 6

Survey and Questionnaire Data

Surveys study various characteristics of individuals from a population. Survey data collection techniques, questionnaire construction to control re accuracy of response rate. Modern surveys grew out of census procedures dating back to the Romans. Today, they are more often directed towards assessing the sentiment of a large population through marketing research, public opinion polls, epidemiological surveys and various national economic, tax and consumption surveys. Surveys are an essential part of managing complex bureaucracies of business, government and public health.

Surveys ask questions to assess constructs such as preferences (e.g., for a tax cut), opinions (e.g., whether drugs are harmful), behavior (e.g., whether trust encourages purchasing), or facts (e.g., family size). The success a survey depends on the representativeness of the sample with respect to a population of interest to the research question. Concerns in surveys range from questionnaire design, execution and interpretation (including follow-up on non-response, and adjustment for sample bias), sample design, data collection instruments, statistical adjustment of data, and data processing, systematic and random survey errors. Cost constraints are imposed by weighing the cost of a data item against the cost of survey error and quality.

Questionnaires are an exceptionally popular survey instrument, and one unfortunately that tends to be too often collected and analyzed carelessly, laxly and incorrectly. Questionnaires have advantages over other survey approaches in that they are cheap and require little effort to implement compared to, for example, verbal or telephone surveys, or interrogative protocols. Questionnaires suffer from many of the same shortcomings as other types of opinion polls.

Questionnaires designed for personal interrogation, such as the widely used Myers-Briggs Type Indicator, strictly limit responses, where respondents can answer either option but must choose only one response. Myers and Briggs, a home-schooled mother and daughter team that popularized typological theories proposed by Carl Gustav Jung in 1921, constructed a relatively ad hoc instrument that remains a bit fuzzy about what question it is actually intended to answer. Such ambiguity is not unusual in questionnaire design. Additionally, most questionnaires suffer significant non-response, especially with mail and online questionnaires; alternatively, people may not care enough to respond truthfully.

Interrogation techniques evolved significantly with the widespread adoption of the polygraph in police and government agencies after the 1930s. Polygraphs, or lie detectors, measure and record physiological indices such as blood pressure, pulse, respiration, and skin conductivity while the subject is asked and answers a series of questions. The theory is that with proper questionnaire design, dishonest answers will produce physiological responses different from truthful answers.

Polygraph testing typically begins with a pre-test interview to gain some preliminary information which will later be used to develop diagnostic questions. This will be followed by a request for respondents to deliberately lie in order to calibrate the 'dishonest' physiological parameters. Then the actual test starts. Some of the questions asked are irrelevant, others are diagnostic questions, and the remainder is the relevant questions. The goal is not just to get responses, but to get true responses, and be able to identify casual or untrue responses. Though polygraph protocols have much to inform survey questions applied elsewhere, these are often ignored because they are considered too much work by many social researchers. Instead, modern marketing, social science, and other survey research disciplines like to provide questionnaires where all questions are relevant, and where all responses are immediately quantified on a Likert scale, so that they can be fed into computer software to generate tables of statistics. The remainder of this chapter investigates the consequences of such an insouciant approach to survey research.

6.1 Rensis Likert's contribution to Social Research

Likert scales are named for Rensis Likert1 who developed them in his PhD thesis, and promoted the 1-5 Likert scale for the remainder of his career. Likert was a founder of the University of Michigan's Institute for Social Research, was the director from its inception in 1946 until 1970, and later founded a consulting firm to promote the Likert scale. The purpose of a Likert (rating) scale is to allow respondents to express both the direction and strength of their opinion about a topic. A Likert item is simply a statement which the respondent is asked to evaluate according to any kind of subjective or objective criteria; generally the level of agreement or disagreement is measured. It is considered symmetric or "balanced" because there are equal amounts of positive and negative positions. The 'distance' between each successive Likert item is traditionally assumed to be equal – i.e., the psychometric distance between 1 and 2 is equidistant to 2 to 3. In terms of good research ethics, an equidistant presentation by the researcher is important; otherwise it will introduce a research bias into the analysis. A good Likert scale will present a symmetry of Likert items about a middle category that have clearly defined linguistic qualifiers for each item. In such symmetric scaling, equidistant attributes will typically be more clearly observed or, at least, inferred. It is when a Likert scale is symmetric and equidistant that it will behave like an interval-level measurement. (U.-D. Reips and Funke 2008) showed that interval-level measurement better achieved by a visual analogue scale. Another perspective applies a polytomous Rasch model to infer that the Likert items are interval level estimates on a continuum, and thus that statements reflect increasing levels of an attitude or trait – e.g. as might be used in grading in educational assessment, and scoring of performances by judges. Often, researchers would prefer respondents to make a definite choice rather than choose neutral or intermediate positions on a scale. For this reason, a scale without a mid-point would be preferable, provided it does not affect the validity or reliability of the responses. Numerous studies have demonstrated that as the number of scale steps is increased, respondents' use of the mid-point category decreases (Friedman et al. 2004) (Friedman, Hastie, and Tibshirani 2010) (Sterne and Smith 2001). (Worcester and Burns 1975) found that grammatically balanced Likert scales are often unbalanced in interpretation; for instance, 'tend to disagree' is not directly opposite 'tend to agree'. Worcester and Burns also concluded that a four point scale without a mid-point appears to push more respondents towards the positive end of the scale. Likert scales may be subject to distortion from at least three causes. Subjects may:

1. avoid using extreme response categories (central tendency bias);
2. agree with statements as presented (acquiescence bias); or
3. try to portray themselves or their organization in a more favorable light (social desirability bias).

Designing a scale with balanced keying (an equal number of positive and negative statements) can obviate the problem of acquiescence bias, since acquiescence on positively keyed items will balance acquiescence on negatively keyed items, but there are no widely accepted solutions to central tendency and social desirability biases. When a Likert scale approximates an interval-level measurement, we can summarize the central tendency of responses using either the median or the mode, with 'spread' measured by standard deviations, quartiles or percentiles. Characteristics of the sample can be obtained from non-parametric tests such as chi-squared test, Mann–Whitney test, Wilcoxon signed-rank test, or Kruskal–Wallis test (Jamieson 2004) (Norman 2010). Likert items are considered symmetric or 'balanced' where there are equal amounts of positive and negative positions. Rensis Likert used five ordered response levels, but seven and even nine levels are common as well. (I. E. Allen and Seaman 2007) concluded that a five or seven point scale may produce slightly higher mean scores relative to the highest possible attainable score, compared to those produced from a ten point scale, and this difference was statistically significant. In terms of the other data characteristics, there was very little difference among the scale formats in terms of variation about the mean, skewness or kurtosis.

6.2 Likert Scales

Likert scales are the most widely used approach to scaling responses in survey research, such that the term is often used interchangeably with rating scale. There is considerable discussion as to the exact meaning of Likert scaling; so much so that this is beyond the scope of this book. The Rasch model is the most intuitive, but not every set of Likert scaled items can be used for Rasch measurement. The data has to be thoroughly checked to fulfill the strict formal axioms of the model (Bond and Fox 2001). Likert scale data can, in principle, be used for obtaining interval level estimates on a continuum by applying the polytomous Rasch model, when data can be obtained that fit this model. In addition, the polytomous Rasch model permits testing of the hypothesis that the statements reflect increasing levels of an attitude or trait, as intended. For example, application of the model often indicates that the neutral category does not represent a level of attitude or trait between the disagree and agree categories. Alternatively, a Likert scale can be considered as a grouped form of a continuous scale. This is important in SEM, since you implicitly treat the variable as if it were continuous for correlational analysis. Likert scales are clearly ordered category scales, as required for correlational work, and the debate among methodologists is whether they can be treated as equal interval scales.

There are two main questions relating to the SEM use of Likert scale datasets. The first one seeks to know the nature of Likert scale and if they can be used for correlation and chi square test. Unfortunately, correlations from two belief distributions – for example (1) belief about whether one is healthy and (2) belief about whether one's career is successful – differ from the correlation of their Likert scale representations. Information is lost in the mapping to a Likert scale. How much information is lost is probably unknowable in most cases (though I address this question from a purely mechanistic standpoint in the next section). There is a body of research (Kühberger 1995) that concludes that people do not generally hold strong, stable and rational beliefs, and that their responses are very much influenced by the way in which decisions are framed. This should serve as a strong caveat in the meticulous design of survey instruments. With a choice of a Fisher Information metric, we can explore the implications of the widespread assumption that survey subject beliefs or other phenomena are Gaussian distributed, but are elicited, measured and analyzed as Likert scaled data. Likert scales have been used since the 1930s to allow respondents to express both the direction and strength of their opinion about a topic (Likert 1932) (Likert 1961) (Murphy and Likert 1938). Thus a Likert item is a statement that the respondent is asked to evaluate according to any kind of subjective or objective criteria; generally the level of agreement or disagreement is measured. Survey researchers often impose various regularity conditions on the metrics implied in the construction of their survey instruments to eliminate biases in observations, and help assure that there is a proper matching of survey results and the analysis (Clarke et al. 2002) (Roberts et al. 2001) (Worcester and Burns 1975). A Likert item in practice is generally considered symmetric or balanced when observations contain equal amounts of positive and negative positions. The 'distance' between each successive Likert item is traditionally assumed to be equal – i.e., the psychometric distance between 1 and 2 is equidistant to 2 to 3. In terms of good research ethics, an equidistant presentation by the researcher is important; otherwise it will introduce a research bias into the analysis. A good Likert scale will present a symmetry of Likert items about a middle category that have clearly defined linguistic qualifiers for each item. In such symmetric scaling, equidistant attributes will typically be more clearly observed or, at least, inferred. It is when a Likert scale is symmetric and equidistant that it will behave like an interval-level measurement. (U. Reips and Funke 2008) showed that interval-level measurement better achieved by a visual analogue scale. Another perspective applies a polytomous Rasch model to infer that the Likert items are interval level estimates on a continuum, and thus that statements reflect increasing levels of an attitude or trait – e.g. as might be used in grading in educational assessment, and scoring of performances by judges.. Any approximation suffers from information loss; specifying the magnitude and nature of that loss, though, can be challenging. Fortunately, information measures of sample adequacy have a long history. These were perhaps best articulated in the 'information criterion' published in (Akaike 1974) using information entropy. The Akaike information criterion (AIC) measures the information lost when a given model is used to describe population characteristics. It describes the tradeoff between bias and variance (accuracy and complexity) of a model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value (minimum information loss); it rewards goodness of fit, while penalizing an increasing number of estimated parameters. The Schwarz criterion (Ludden, Beal, and Sheiner 1994) is closely related to AIC, and is sometimes called the Bayesian information criterion. Ideally, responses to survey questions should yield discrete measurements that are dispersed and balanced – this maximizes the information contained in responses. Researchers would like respondents to make a definite choice rather than choose neutral or intermediate positions on a scale. Unfortunately, cultural, presentation and subject matter idiosyncrasies can effectively sabotage this objective (for example see (Dietz, Bickel, and Scheffer 2007) and (Lee et al. 2002)). (Lee et al. 2002) point out that Asian survey responses tend to be more closely compressed around the central point than Western responses; superficially, this suggests that Asian surveys may actually yield less information (dispersion) than Western surveys. To improve responses, some researchers suggest that a Likert scale without a mid-point would be preferable, provided it does not affect the validity or reliability of the responses. (Cox III 1980) (Friedman and Amoo 1999) (Friedman, Wilamowsky, and Friedman 1981) (Komorita and Graham 1965) (Matell and Jacoby 1972) (Wildt and Mazis 1978) have all demonstrated that as the number of scale steps is increased, respondents' use of the mid-point category decreases. Additionally (Worcester and Burns 1975) (Sparks et al. 2006) (Dawes and Cresswell 2012) (Meshkati et al. 1995) (Chan 1991) have found that grammatically balanced Likert scales are often unbalanced in interpretation; for instance, 'tend to disagree' is not directly opposite 'tend to agree'. Worcester and Burns also concluded that a four point scale without a mid-point appears to push more respondents towards the positive end of the scale.

The previously cited research studies, in summary, conclude that Likert scales are subject to distortion from at least three causes. Subjects may:

1. Avoid using extreme response categories (central tendency bias);
2. Agree with statements as presented (acquiescence bias); or
3. Try to portray themselves or their organization in a more favorable light (social desirability bias).

Designing a balanced Likert scale (with an equal number of positive and negative statements) can obviate the problem of acquiescence bias, since acquiescence on positively keyed items will balance acquiescence on negatively keyed items, but there are no widely accepted solutions to central tendency and social desirability biases. Likert items are considered symmetric or 'balanced' where there are equal amounts of positive and negative positions.

The number of possible responses may matter as well. Likert used five ordered response levels, but seven and even nine levels are common as well. (I. E. Allen and Seaman 2007) concluded that a five or seven point scale may produce slightly higher mean scores relative to the highest possible attainable score, compared to those produced from a ten point scale, and this difference was statistically significant. In terms of the other data characteristics, there was very little difference among the scale formats in terms of variation about the mean, skewness or kurtosis. From another perspective, a Likert scale can be considered as a grouped form of a continuous scale. This is important in path analysis, since you implicitly treat the variable as if it were continuous for correlational analysis. Likert scales are clearly ordered category scales, as required for correlational work, but the debate among methodologists is whether they can be treated as equal interval scales. When a Likert scale approximates an interval-level measurement, we can summarize the central tendency of responses using either the median or the mode, with 'dispersion' measured by standard deviations, quartiles or percentiles. Characteristics of the sample can be obtained from non-parametric tests such as chi-squared test, Mann–Whitney test, Wilcoxon signed-rank test, or Kruskal–Wallis test (Jamieson 2004) (Norman 2010). Likert mappings may also be analyzed with respect to their resolution or granularity of measurement. Clearly a nine-point scale mapping has more resolution (or finer granularity) than a three point one. Measurement in research consists in assigning numbers to entities otherwise called concepts in compliance with a set of rules. These concepts may be 'physical', 'psychological' and 'social'. The concept length is physical. But the question remains, 'if I report length as 6 feet in a case, what exactly does that mean? Even with physical scales, there is an implied granularity; if I say that something is 6 feet long, this implies less precision than length of 183 centimeters. In scientific pursuits, finer granularities can be pursued to almost unimaginable levels – for example, the international standard for length, adopted in 1960, is derived from the 2p10-5d5 radiation wavelength of the noble gas Krypton-86. The influence of choice of measuring stick on the results of modeling is responsible for phenomena such as Benford's Law (Benford 1938) (Hill 1995) and fractal scaling (Basmann 1963b) (Mandelbrot 1982). The assumption of Gaussian distribution of opinions or beliefs is common in the analysis of survey research, as mentioned previously. The assumption tends to be applied in the analysis stage, rather than in the design of the survey instrument. The question of whether underlying beliefs are continuous or discrete, distributed one way or another doesn't tend to come up in the design of Likert scaled surveys, because there a few conventions that could use this information to improve the survey design. Nonetheless, the current research will argue that it matters in assessing the informativeness of Likert scaled data, which in turn can have a large impact on significance, power and other statistics reported from the research. Information clearly is lost in the mapping of beliefs to a Likert scale; how much information is lost is probably unknowable in practice. But the loss in information from that that would exist if our modeling assumptions (e.g., Gaussian beliefs) were actually true can be assessed.

6.3 How much information is lost in Likert responses?

Beliefs are inherently continuous, ambiguous and changeable. Our ability to hold beliefs about abstractions evolved to help us manage the complex and dangerous environment of prehistoric African savannas; they were did not evolve to yield accurate marketing studies. Likert scales, as useful as they may be, consequently suffer from problems involving (1) informativeness, (2) bias, and (3) dispersion in Likert representations of survey subject beliefs.

Proper instrument design requires a standardization of Likert responses so that ideally one standard deviation of the actual distribution of beliefs will shift the Likert score one point higher – this is comparable to the process of keeping the survey instrument 'on scale' measuring beliefs in similar units to the subjects normal conventions. In addition, the mode of subject beliefs (i.e., what the largest number of people believe or agree upon) is presumed to center somewhere in the range 2 through 4 of the 5-point scale, with all other values being

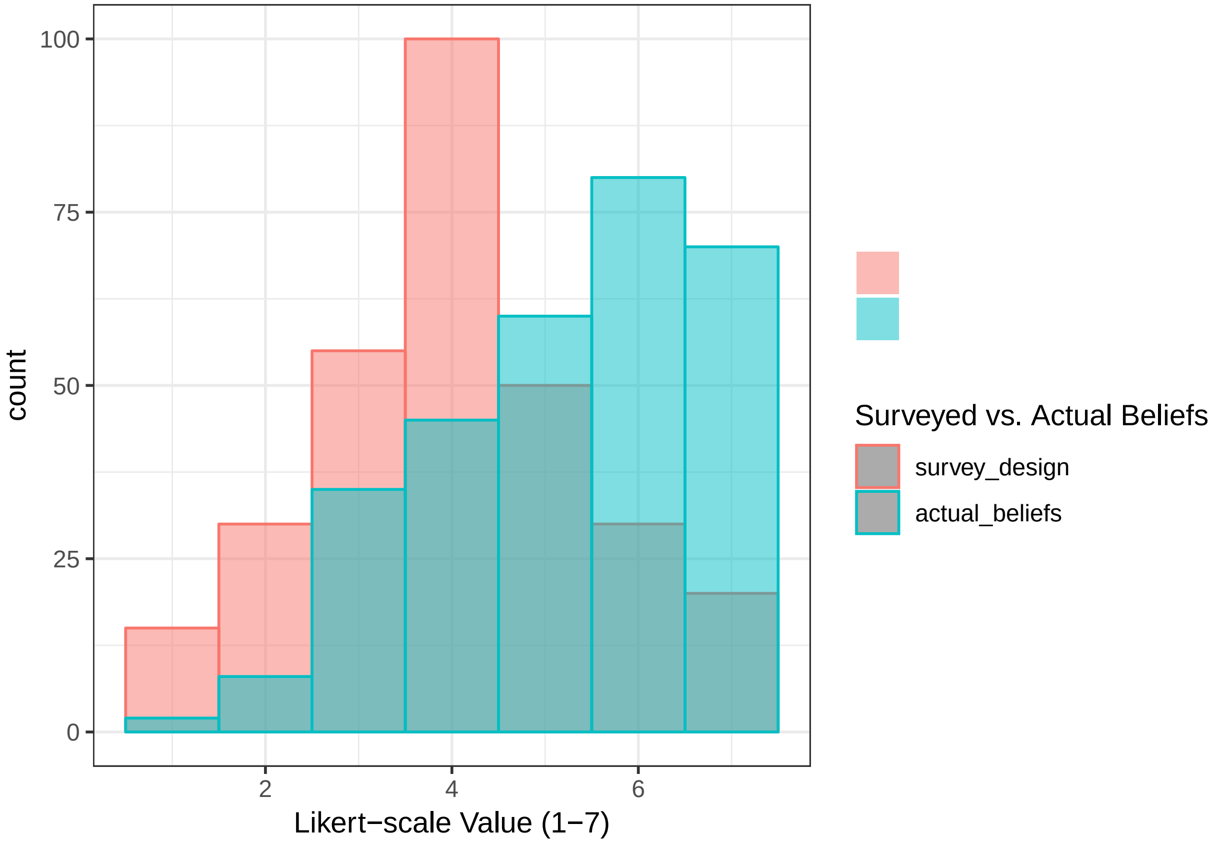

Figure 6.1: (#fig:likert_polite)Positive Survey Bias

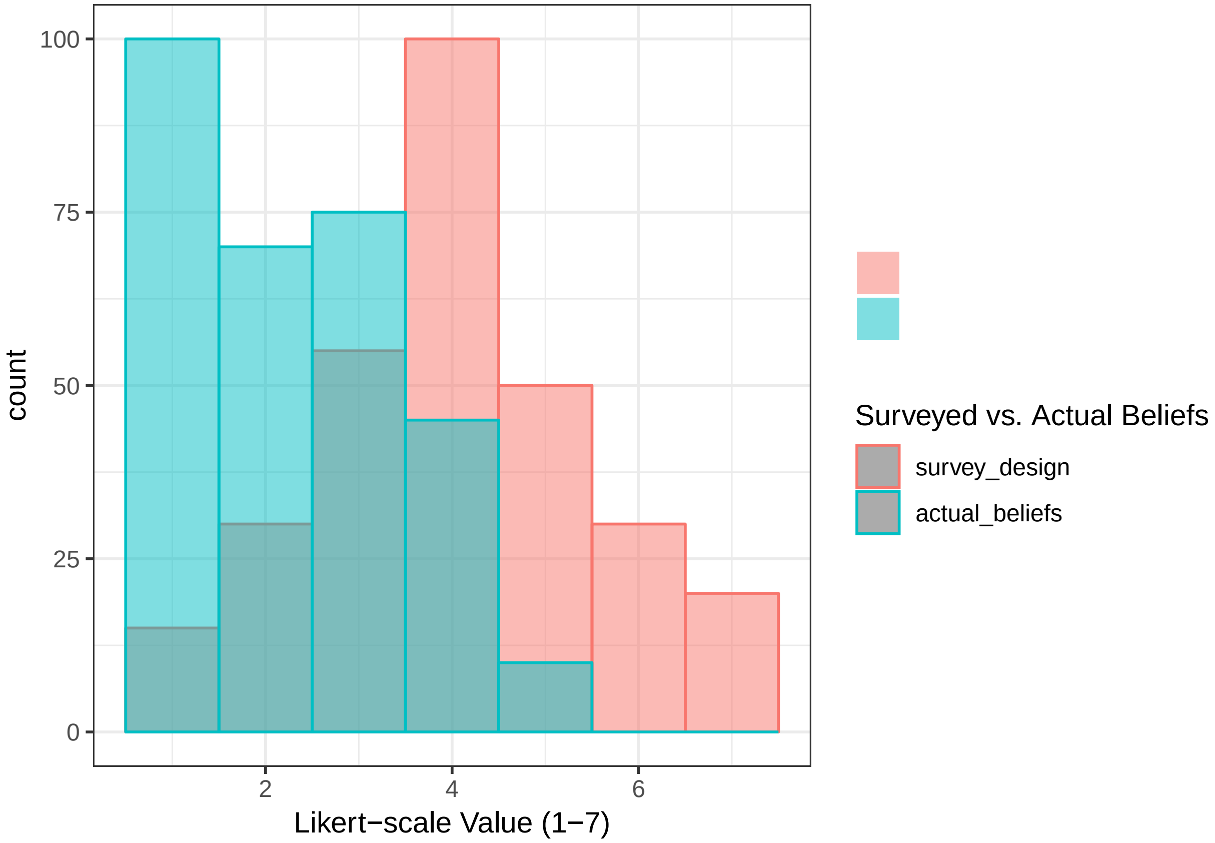

Figure 6.2: (#fig:likert_rude)Negative Survey Bias

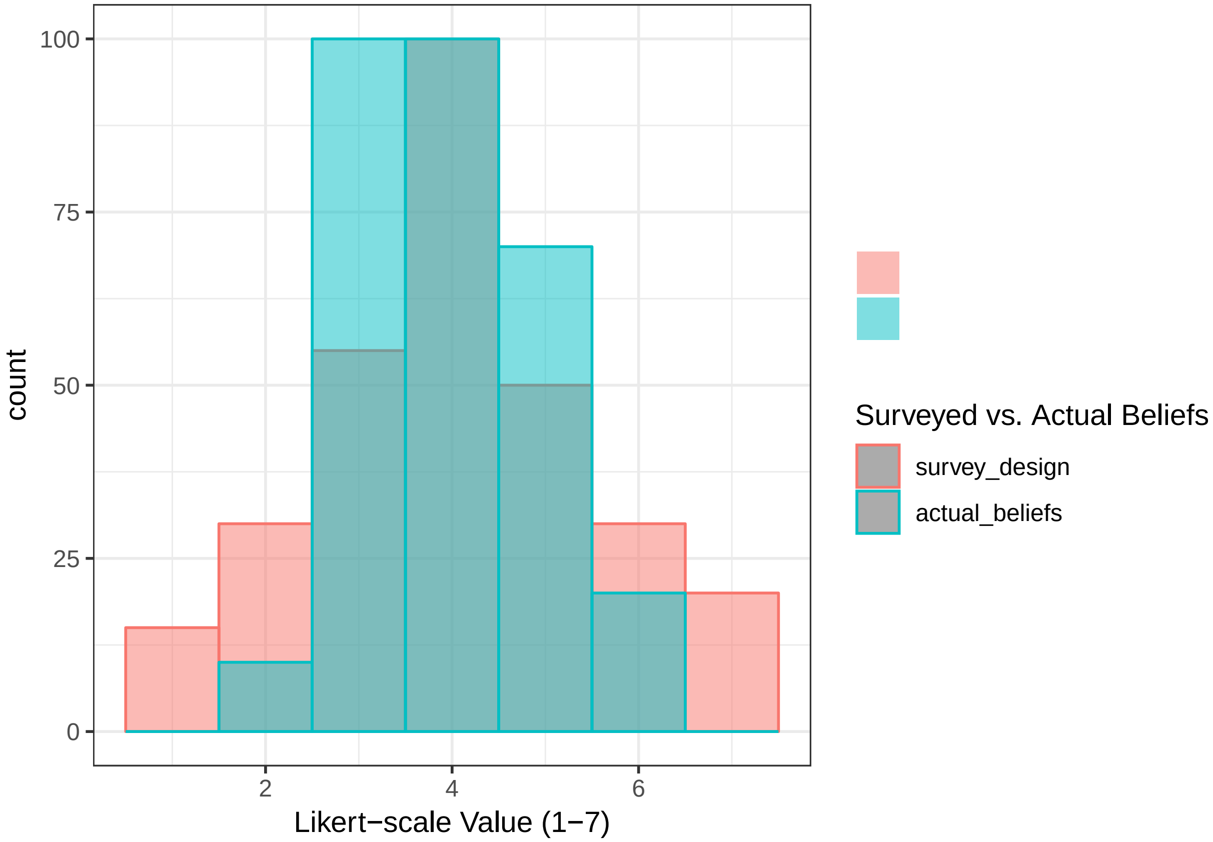

Figure 6.3: (#fig:likert__asian)Asian Survey Bias

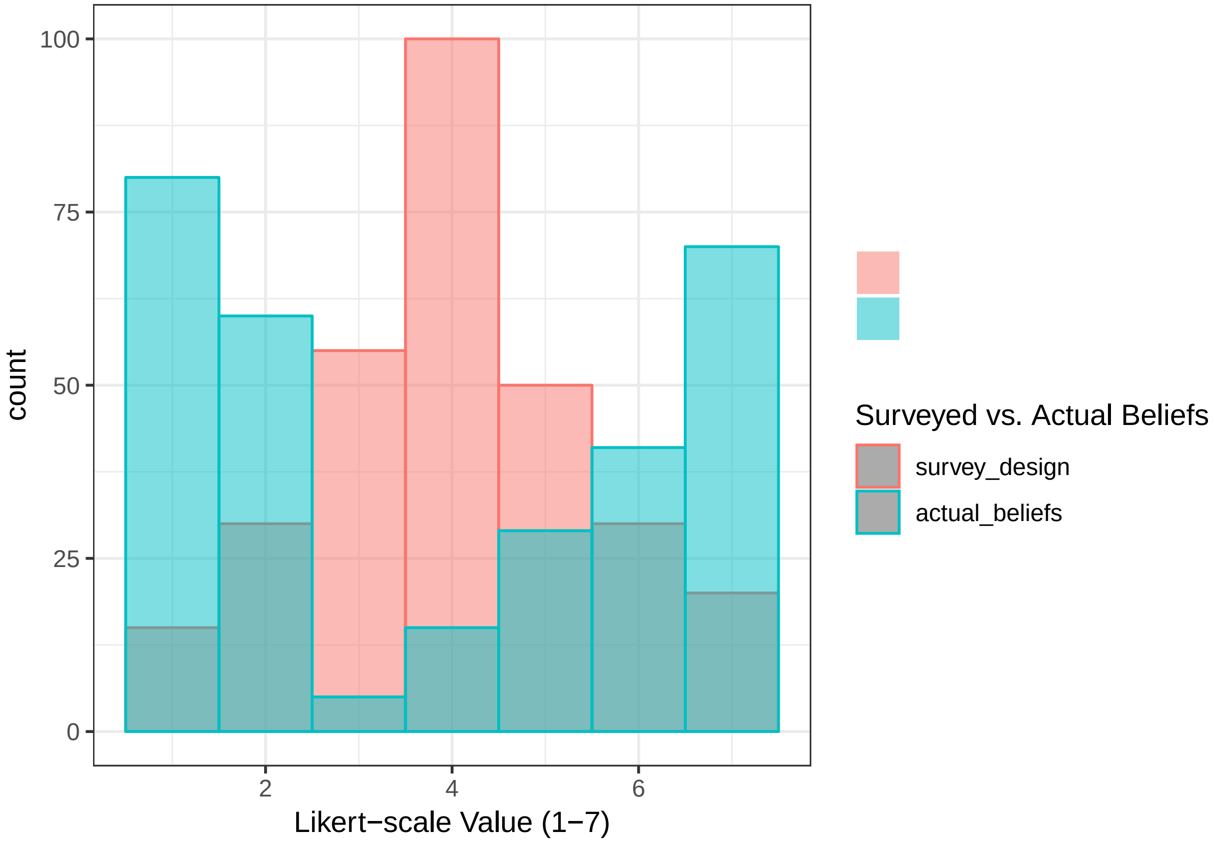

Figure 6.4: (#fig:likert_american)American Survey Bias

the ‘extremes’ – response ‘1’ or response ‘5’. This is more or less what survey researchers aspire to, where the level of agreement or disagreement is measured (i.e. is ‘on scale’) and the scaling is considered symmetric or ‘balanced’ because there are equal amounts of positive and negative positions (A. C. Burns & Bush, 2005; A. Burns & Bush, 2000). Most of the weight of the Gaussian belief distribution should lie within the Likert range 2 through 4 of the 5-point scale. Survey researchers can credibly move the range around, but probably should not try to alter the subject beliefs if they are trying to conduct an objective survey. Weaknesses in data can be effectively addressed by increasing the sample size. This works for multicolinear data, non-Gaussian data, and for Likert data as well. But since data collection is costly, it is desirable to increase sample size as little as possible. The path analysis literature is surprisingly vague on how much of an increase is needed. (Jöreskog, 1971a, 1971b) suggest increases of two orders of magnitude, but without offering causes or mitigating factors. If we assume that survey costs increase commensurately with sample size, then for most projects two orders of magnitude is likely to be prohibitive. For example, in the path analysis approaches implemented in LISREL and AMOS software, for reasonably large samples, when the number of Likert categories is 4 or higher and skew and kurtosis are within normal limits, use of maximum likelihood is justified. In other cases some researchers use weighted least squares based on polychoric correlation. (Jöreskog and Goldberger 1975) investigated Monte Carlo simulation finding that phi, Spearman rank correlation, and Kendall tau-b correlation performed poorly whereas tetrachoric correlation with ordinal data such as Likert scaled data was robust and yielded better fit.

6.4 The Information Content of Items Measured on a Likert Scale

Assume a survey collects n independent Likert-scaled observations for each survey questions. Let the Likert scale represents a polytomous Rasch model (with say 5, 7 or 9 divisions). We can take the perspective of a polytomous Rasch model, assuming that the responses to the survey map an underlying Gaussian belief distribution to a Likert item across the population of subjects surveyed for a particular question on the survey. Ideally, survey responses will yield an equidistant scaling of Likert items; for example one standard deviation of the actual distribution of beliefs will shift the Likert score one point higher. In addition, let the mean value of the mean of beliefs is presumed to center somewhere in the range 2 through 6 of the 7-point scale, with all other values being the ‘extremes’ – response ‘1’ or response ‘7’. This is more or less what marketing researchers aspire to, where the level of agreement or disagreement is measured (i.e. is ‘on scale’) and the scaling is considered symmetric or ‘balanced’ because there are equal amounts of positive and negative positions (Burns and Bush 2005). Most of the weight of the Gaussian belief distribution should lie within the Likert support (we presumably can move the Likert support around, but we probably should not try to alter the subject beliefs if we are running an objective survey). Let $F(. \mid \mu, \sigma)$ and $f(. \mid \mu, \sigma)$ be cumulative distribution function and probability distribution function respectively of the underlying belief distribution. Presume we use a metric scale that sets

σ (or alternately that the Likert 'bin' partitions are spaced σ units apart). A particular bin i is filled with probability of p and not filled with probability $1-p$; let n independent survey questions result in that bin being filled $i\theta$ times, and not filled $n-i\theta$ times. If B_i is a logical variable that indicates whether the bin of the Likert item was chosen, then all possible outcomes for the Likert item can be represented $B_1 \cup B_2 \cup ...B_{k-1} = \cup_{i=1}^{k-1} B_i$ since if none of the first $k-1$ bins were chosen, then the k^{th} bin must have been chosen. Let the Fisher information in the i^{th} bin of a sample of n Likert items be I_{B_i}. Since B_i is a logical variable, it can be perceived as a Bernoulli trial ˘ a random variable with two possible outcomes, "success" with probability of p and "failure", with probability of $1-p$. The Fisher information contained in a sample of n independent Bernoulli trials for B_i where there are m successes, and where there are $n-m$ failures is:

$$I_{B_i}(p_i) = -E_{p_i}[\frac{\partial^2}{\partial^2 p_i} ln[f(m \mid p_i)]] = \frac{n}{p_i(1-p_i)}$$

which is the reciprocal of the variance of the mean number of successes in n Bernoulli trials, as expected. The Fisher information contained in a sample of n independent Bernoulli trials for all possible outcomes for n Likert items $\cup_{i=1}^{k-1} B_i$ is:

$$I_{\cup_{i=1}^{k-1} B_i} = \sum_{i=1}^{k-1} (\frac{n}{p_i(1-p_i)}).$$

Compare this to the Fisher information in a sample of n observations from a Gaussian $N(\mu, \sigma^2)$ belief distribution, which is estimated $I_n = \frac{n}{\sigma^2}$ and is independent of the location parameter μ. From these formulas, we can construct a Likert 'penalty' ω that is the ratio of information content in these two different mappings from the survey sample:

$$\omega \triangleq \frac{\frac{n}{\sigma^2}}{\sum_{i=1}^{k-1} (\frac{n}{p_i(1-p_i)})} = \frac{1}{\sigma^2 \sum_{i=1}^{k-1} (\frac{1}{p_i(1-p_i)})}$$

This ratio provides a guide to sample sizes for Likert scaled versus more refined continuous metrics. The lower bound on a Likert-scaled sample will need to be ω time as large as an ideal measurement instrument that is able to perfectly measure the subjects' beliefs. Such an ideal is in practice not available, but our calculations demonstrate that there is substantial information loss when using Likert-scale surveys. Indeed, proper survey protocols are important to assure that these information losses do not make any survey too unreliable. There are three things that should be noted concerning multiplier for sample size estimates for processing Likert data when an assumption of Gaussian data has been made in the data analysis:

1. any difference of the actual sample standard deviation from the equidistant scale of the Likert items requires larger sample sizes; but the minimum sample size for any Likert mapped data set will be at 100 times as large as that that would be required if you had all of the original information in the Gaussian distribution of beliefs. The information loss from using Likert scaling is at least two orders of magnitude – the increase in sample size is at least two orders of magnitude.
2. the sample is most informative when location of the Gaussian mean coincides with the central Likert bin. This emphasizes the importance of 'balanced' designs for the Likert scaling in the survey instrument.
3. information in the underlying belief distribution, which has a support, does not depend

on the mean of an assumed underlying Gaussian distribution of data. The Likert mapping information content does depend on the mean and is sensitive to the Likert scale being ‘balanced’ – this is controlled in the survey design.

6.5 Affective Technologies to Enhance the Information Content of Likert-scaled Surveys

Primitive affective technologies were employed for a variety of purposes, from the spiritual to the forensic, since prehistory. For example, crime suspects in ancient China were told to hold a handful of rice in their mouths. If the rice stayed dry, they were absolved of the crime. Another example is firewalking, which was a recognized spiritual objective in many beliefs. Similarly, in other societies, ‘trial by ordeal’ absolved possible criminals if they were not burned by hot irons. And later, interrogators scrutinized the involuntary facial micro-movements to determine whether a person was lying. Affective technologies for eliciting accurate subject responses have developed throughout the 20th century, beginning with (Jung 1919) and later in forensics (Tao and Tan 2005)(Marston 1938) and most recently in robotics and game technology (Picard 2010).

One promising affective technology for enhancing the information content of surveys, electrodermal response (EDR), was a laboratory curiosity long before it was applied in a practical setting. The use of the electrical measurement to detect emotion dates to the time of its invention by Italian physiologist Luigi Galvani in his paper on animal electricity in 1791. The seminal study of psychosomatic bioelectric effects appeared in the 1870 s in the study of basal skin resistance and stimulated skin resistance in connection with psychological state (Féré 1899). It was suggested that this work could be applied to the fields of criminology and hypnotherapy among other uses (James 1884). Independent studies (see Neumann and Blanton 1970) found evidence of correlation between electricity and human emotions. So-called electrodermal studies were seen as one way to read human emotional responses through observations of physiological changes, such as sweating and involuntary muscle contractions.

After nearly a century of laboratory studies by the early innovators with the technologies, EDR response methods were applied in a repeatable, scientific methodology,which led to practical applications and the use by early adopters (Marston 1938). This provided a basis for wider use of EDR based approaches and also made EDR technologies popular with practicing psychotherapists. Others described clinical techniques for EDR measurement, while words were read to the subject from a prepared list. If a word on this list was emotionally charged, there was a change in body resistance, and this caused a reading. This allowed the EDR technologies to become reliable sources of data for psychological questionnaires, foreshadowing their applications in forensics, marketing and surveys.

Early EDR measurement devices were not simple to use. They were fundamentally Wheatstone bridge null detectors that ran a small current through the subject’s palms or soles of his feet, while varying a resistor until a moving arm galvanometer read zero. They also lacked amplification: the electron tube had not yet been commercialized. This tended to make them inaccurate and inconsistent in their readings. This null configuration was useful in the early days, when temperature and humidity might make component values drift. When they drifted in the same direction, rough accuracy would be maintained. With the commercialization of vacuum tubes in the beginning of the last century, electrical engineers began experimenting with amplifying the Wheatstone bridge-galvanometer null detector setup that existed

Amplification made EDR technologies much less troublesome to use, and made electrodermal measurements available to the majority of adopters by 1930. Polygraphs, as machines for measuring affective data streams were called, became widespread and affordable in the 1920s and 1930s. Both methodology and user interfaces were studied and standardized by Marston (Marston 1938), who wrote on the theory and use of polygraphs in law and marketing. Marston was a polymath who is most famous for originating the comic book character Wonder Woman, whose lasso was modeled after the pneumatic band used in Marston's lie detectors.

Marston famously applied affective technologies in advertising for the Gillette company, to claim that:"My study enables me to state lately that Gillette Blades are far superior in every respect to competitive blades tested."Marston's public stature encouraged innovations in criminal law enforcement adopted by the Chicago Police Department, which promoted the application of new interrogation techniques based on affective technologies— particularly EDR technologies—and other law enforcement agencies followed Chicago's lead. The development of the polygraph integrated EDR measurement devices with respiration, blood pressure and perhaps temperature, and other physiological responses to stress. The layering of affective measurement devices reflected the limitations in any particular physiological measurement. Affective forensics were so effective, however, that legal and commercial use of lie detectors grew rapidly after World War II, with over 100 companies producing machines to read affective data streams by the early 1950s.

These technologies are now mature, and EDR measurement devices are widely adopted, though interpretation is still controversial. EDR technologies appear in other contexts as well: biofeedback and protocols designed to help control epileptic seizures (Ramaratnam, Baker, and Goldstein 2008); relaxation and analysis within various spiritual groups (Critchley et al. 2002) (Critchley 2002); as a controller for video games and other computer interfaces (Headon and Curwen 2002); and in forensic interviews.

EDR measurement technologies have grown considerably more reliable and affordable with the development of digital electronics. They are now popular with hobbyists, and many low-cost designs have been posted on the Internet. This, in turn, has made adoption of popular innovations in consumer electronic devices and games. Low-cost analog to digital converter boards are widely available now too, and can be used to provide EDR measurement device resistance measurement in digital form.

EDR is best measured on the palms of the hands. This produces the greatest detail and fastest response in changes in skin resistance. There are two types of electrodermal response measures based on the resistance between two electrodes placed in the palmar region of the hand. The first type of measurement is referred to as exosomatic, since the current on

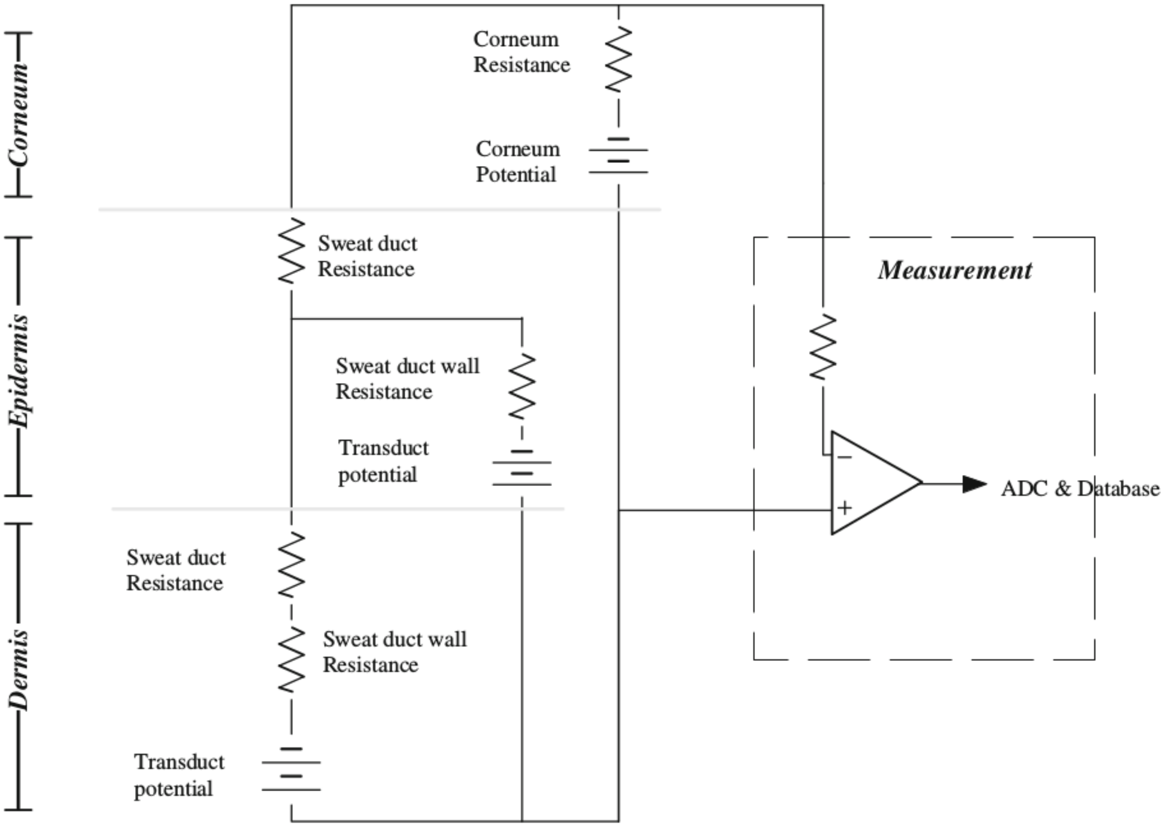

Figure 6.5: (#fig:aff_skin)Conceptual schematic of pathways responsible for electrodermal response

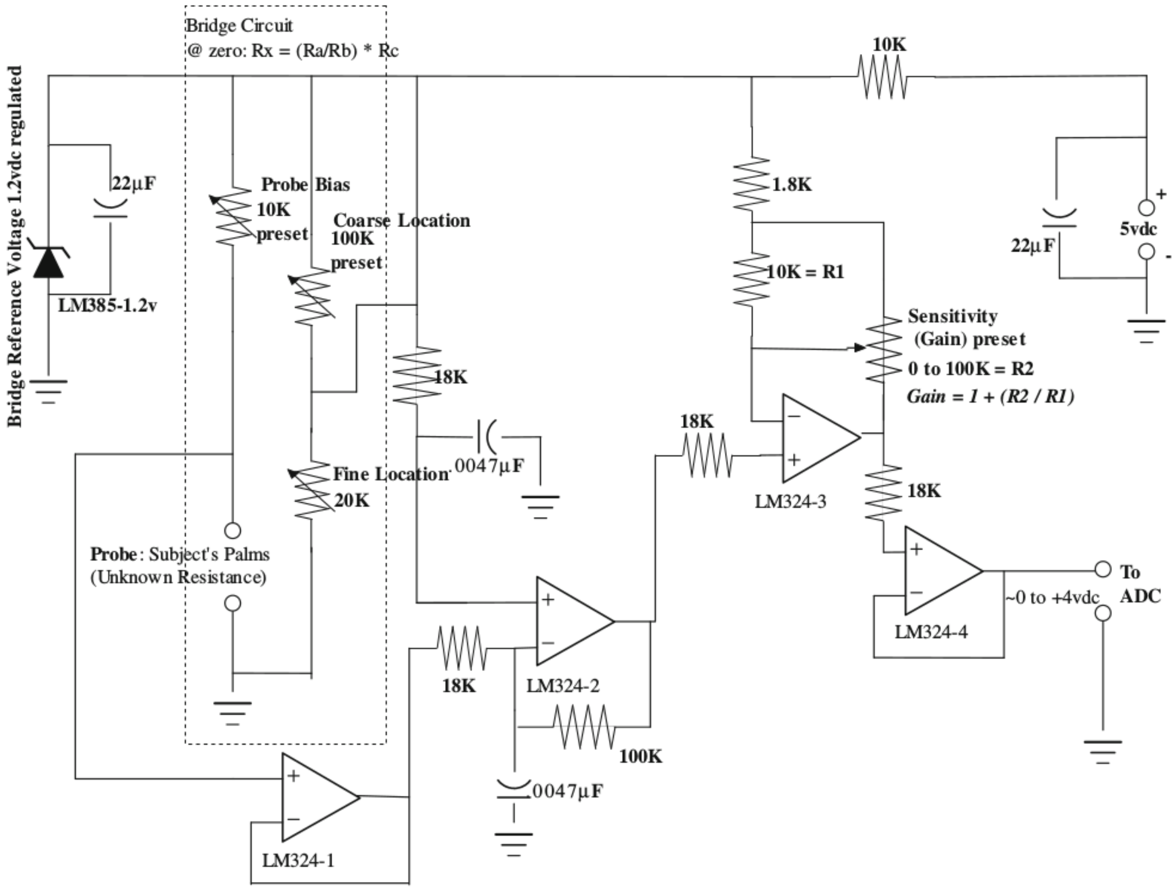

Figure 6.6: (#fig:aff_circuit)Simple Wheatstone bridge circuit for measuring EDR

which the measurement is based is introduced from the outside. The second type, which is less commonly used, is called endosomatic, since the source of voltage is internal. EDR measurement places a regulated voltage with constant current and constant voltage of around 2.5 volts DC using dry contact electrodes. Neural responses are typically classified into tonic, or regular spiking, responses, and phasic or bursting responses, with approximately 0–5 Hz for tonic measurements, and 0.03–5 Hz being adequate for phasic measurements. Recommendations for electrodermal measurements, drawn up by a committee, have been published in Psychophysiology (see (Greco et al. 2016) (Sequeira et al. 2009) (Greco, Valenza, and Scilingo 2016) (Fowles et al. 1981)).

The application of EDR technologies to enhance survey research have been studied in (Westland 2011a) and (Westland 2011b) and in particular how much relevant information about a specific topic are we able to gain from a typical application of EDR data acquisition in survey questionnaire-based research.

The research approach follows polygraph protocols that demand create a survey response environment that is better-controlled than standard survey protocols. In particular, polygraph survey protocols, in addition to the primary relevant statements on the survey, use irrelevant, truthfulness and sacrificial statements The term sacrificial statement arose in the protocols designed by Marsden for using the polygraph, and have been a standard protocol for analysis of affective data ever since. Irrelevant and truthfulness statements attempt to benchmark the survey subject's mood, cooperativeness and seriousness about the survey. The research assumes that there will have to be certainly statements that allow the survey analysis to be calibrated to account for natural variations in emotional responses between subjects.

The following is a summary of these statement types, and their purpose in the survey. The letter codes to the left of each statement : - Irrelevant statements include the very first statement as well as two others in the center of the survey instrument. Irrelevant statements are designed to identify the subject who are obviously "gaming" the survey, and who are not serious. - Truthfulness statements involve behavior that the majority of subjects have been involved in. Thus, when the subject answers 1–3 = disagree, this is probably a lie. Comparison statements are designed to identify the subject who is 'gaming' the survey. - Sacrificial statements always immediately precede the first of the set of three statements eliciting information on a particular topic. It is worded so that the subject answer is 5–7 = agree. These are designed to absorb the initial response to a relevant issue and to set the context so that subsequent statements on the same issue are recognized, and elicit consistent responses. They are not included in the evaluation/ conclusions. - Primary relevant statements are those that are information bearing, and provide information used in drawing conclusions from the study.

For example, with a survey coding with a Likert scale which runs from 0 = strongly disagree to 7=strongly agree and where codes in parentheses to the left of each statement have a letter code (type of statement):

```
- (I) I am now in Chicago
- (S) I am a conservative
- (P) Only the wealthiest Americans are happy
- (P) We must spend to promote economic growth
- (T) I have lied to a public official
- (S) I believe in minimal government
- (P) Our top national security priority is waging wars
- (T) I have stolen office supplies
```

The development and standardization of affective-verbal survey methods should be a priority in research disciplines if surveys are to retain their credibility in an era of massive cloud based data resources. The integration of the affective response, as measured by a polygraph, into the scoring of particular information received from the subject is often performed in an ad hoc manner, dependent on the skill and experience of the person administering the test. (Westland 2011a) provided mathematical models that can be used to integrate the verbal (Likert-scale) information with affective (polygraph) information in a formal, replicable, standardized fashion. (Westland 2011a) and (Westland 2011b) showed that the integrated responses convey more information than Likert-scaled surveys alone. Survey techniques that validate Likert-scale responses with irrelevant, truthfulness and sacrificial statements are substantially less noisy, and more reliable than naked survey questionnaire responses.

6.6 Known Unknowns: What is a Latent Variable?

What do we really know about our unmeasurable, so-called 'latent' variables. We have defined them as constructs that we think are real, but which cannot be directly measured. In a social context, these may be abstractions like trust, intent, happiness and so forth. In other cases latent variables are cultural – in some cultures, trust may only apply to family; in others, it might be a distinguishing feature of national culture. Latent variables might not even be real – rather they could be shared perceptions, possibly even fantasies. For example, one survey concluded that 77% of adult Americans believe in angels (Johnson, 2011). Thus 'angels' appear to be appropriate latent constructs. If you are not comfortable trying to find indicator variables for a structural model of latent variables germane to angels, then consider the results of another poll – 84% of children believe in Santa Claus. We could construct some structural relationships between Santa's reindeer, aerodynamics, and overall mass – all unobservable – and build a measurement model on subjects' perceptions of Santa. Given the high rate of belief in Santa's existence, we are likely to experience high response to any survey we would construct. In both cases, the measured variables would be perceptions, since it would be difficult to acquire physical evidence of either angels or Santa. Whatever the basis in reality is for one's structural model, its implementation will involve a linear combination of measured factors. We see this elsewhere in statistics, in contrasts: linear combinations of two or more factor level means whose coefficients add up to zero. In SEM, clustering of measured factors around latent variables is decided a priori by the researcher as a part of model specification, perhaps based on pretests and principal components analysis.

From a practical standpoint, the structural, or inner model is merely an artifact that identifies our model as a structural equation model. We could just as easily substitute a linear combination of measured factors everywhere that a latent variable appears in the model. And thus from a practical standpoint, any latent variable structural equation model has an equivalent linear model that is constructed entirely of observed variables.

Chapter 7

Research Structure and Paradigms

John Ioannidis, a highly respected medical researcher, has a serious bone to pick – with the modern practitioners of the Galilean hypothetico-deductive model-data duality discussed in Chapter 6. Ioannidis 2005 paper (J.P.A. Ioannidis 2005) provocatively titled "Why Most Published Research Findings Are False," has been the most downloaded technical paper in PLoS Medicine and one of the single most cited and downloaded papers (J.P.A. Ioannidis 2005). In it, Ioannidis analyzed "49 of the most highly regarded research findings in medicine over the previous 13 years" comparing them with data from subsequent studies with larger sample sizes. His findings: 7 (16%) of the original studies were contradicted, 7 (16%) the effects were smaller than in the initial study, 20 (44%) were replicated and 11 (24%) of the studies remained largely unchallenged (Freedman 2010) (JP Ioannidis 2005) (McCarthy et al. 2008) (Liberati et al. 2009).

Ioannidis' 'most research findings are false' assertion was not hyperbole; indeed only 44% of these highly regarded findings could be replicated. His research was surprising and influential, and resulted in subsequent changes in the conduct of U.S. clinical trials. Ioannidis called this failure to replicate findings the "Proteus phenomenon." Weak research is often driven by an incentive to publish quickly, for fame, reputation, patent rights or ability to publish results at all. Research priority is the credit given to the individual or group of individuals who first made a discovery or propose a theory. Priority debates have defined the form an context of modern science, yet as Stephen Jay Gould once remarked "debates about the priority of ideas are usually among the most misdirected in the history of science." (Gould 1977) Priority has been at the center of Western research traditions for four centuries. It is the primary reason that research journals exist. The early research journal Philosophical Transactions of the Royal Society was founded in the 17th century at a time that scientists did not publish; rather they competed in contests for employment. At that time, the act of publishing academic inquiry was similar to distribution of open-source software today: difficult to justify because of the lack of financial incentives. Yet it was highly effective in resolving priority disputes. Studies found that 92% of cases of simultaneous discovery in the 17th century ended in priority dispute; this dropped to 72% in the 18th century, 59% by the latter half of the 19th century, and 33% by the first half of the 20th century (Merton 1968). That is not to say that publishers necessarily got priority right. A cynical, but widely accepted view is called

Stigler's law of eponymy: no scientific discovery is named after its original discoverer. Stigler drolly named the sociologist Robert K. Merton as the discoverer of Stigler's law to avoid contradiction (Stigler 1980) referring to Merton's 'Matthew Effect' (Merton and Merton 1968) (Merton 1988) (Merton 1995) in which the rich get richer, the powerful more powerful, and the poor more destitute. The Proteus phenomenon is less of an issue in the social sciences, only because it is virtually impossible to replicate the quasi-experiments that are the norm in the social sciences. Consequently, social science research findings are not subject to the same intense (and well-funded) scrutiny of medicine. That doesn't mean that they don't suffer from their own Proteus phenomenon. Since problems are unlikely to be detected after publication, control over social science research protocols must happen earlier in the process – at the time the research is designed. This chapter addresses the problems and potential controls over social sciences' own Proteus phenomenon.

7.1 The Quest for Truth

Statisticians like to think of their craft as a game, pitting them against nature, which keeps secret the true state of thing. Answering questions gives statisticians insight into the 'true state of nature' – into the real world. Philosophers have been seeking the truth throughout much of the World's history. Some if the salient theories to arise from this quest have been:

1. Correspondence theories claim that true beliefs and true statements correspond to the actual state of affairs

2. Coherence theories requires a proper fit of elements within a whole system as a basis for asserting truth.

3. Social constructivism holds that truth is constructed by social processes, is historically and culturally specific, and that it is in part shaped through the power struggles within a community. Constructivism views all of our knowledge as 'constructed,' because it does not reflect any external 'transcendent' realities

4. Consensus theory holds that truth is whatever is agreed upon, or in some versions, might come to be agreed upon, by some specified group.

5. Pragmatic theory articulated by the psychologist William James (Merton and Merton 1968): 'the 'true' is only the expedient in our way of thinking, just as the 'right' is only the expedient in our way of behaving.'

6. Minimalist (deflationary) theories reject the thesis that the concept or term truth refers to a real property of sentences or propositions

The exact meaning of 'Truth' is open to wide interpretation requiring strong arguments. This may require more energy than statisticians may be willing to give up to the philosophical portion of their projects. Most are satisfied with G.E.P. Box's dictum that 'All models are wrong, but some are useful' (Box and Draper 2007). This chapter takes Box's (and indirectly James') 'pragmatic' approach to the truth.

7.2 Research Questions

No matter what the topic, the most important decision facing the researcher is choice of research question. This choice determines the data collected, the method of analysis used, and the ultimate meaning and utility of any answer that research might provide.

Data acquisition often is the costliest part of any project, and there is a natural tendency to look for questions that that data can answer. This is understandable, and it can work if the researcher is honest in seeking a question to answer. Datasets are often quite limited by design in the amount and quality of information they contain, simply because of the cost trade-off. Exploratory data analysis is directed towards finding out what information is contained in a database – and thus what questions can be answered. Research questions are dependent on dataset information– you cannot ask a research question of a database that it is unprepared to answer (no matter how much you torture the data).

7.3 Models

A model is a theoretical construct that represents something, with a set of variables and a set of logical and quantifiable relationships between them. Models are constructed to enable reasoning within an idealized logical framework about these processes; they are an important component of scientific inference and deduction When we use the term ‘idealized,’ we mean that the model can make explicit assumptions that are known to be false (or incomplete) in some detail. Such assumptions may be justified on the grounds that they simplify the model, while at the same time, allowing the production of acceptably accurate solutions. Another perspective would be that make these false, incomplete, simplifying assumptions knowing that they will product errors, and the trade off is with the quantifiable inaccuracy, resolution or granularity of the errors in our conclusions. Tweaking these assumptions in subsequent research facilitates a stepping-stone approach to research, just as we might pause to catch our balance at each stepping stone in crossing a river (rather than trying to cross in one go). This latter perspective assumes that a particular research project is embedded in a more comprehensive research program. Research programs are inclined to adopt the sort of new venture options approaches that we see in investment and industry – and for a similar end: to use scarce research time and funding in the most efficient way possible.

Venture options approaches parcel work out in increasing quantities of time and research funding. They adopt a tiered series of projects, often involving three steps of funding:

1. Proof of concept or pretest;
2. Limited testing; and
3. Full set of tests.

There are likely to be qualitative differences in these tests as well. They will fall into three categories:

1. Positioning options; these tie down the most useful and efficient set of assumptions for the models used in the full set of tests.

2. Scouting options; Former Secretary of Defense Donald Rumsfeld once famously remarked that there are ‘known unknowns, and unknown unknowns.’ Scouting options are designed to bring the latter ‘unknowns’ into the realm of the ‘knowns’; and
3. Stepping-stone options: this provides a formal plan for the stepping-stone approach to model specification, allowing us to collect small data sets to test the applicability of assumptions while we are still deciding the final form of the model. Exploratory analysis statistical techniques are typically very useful in analysis with a stepping-stone option approach.

7.4 Theory Building and Hypotheses

Differing objectives and traditions are likely to be associated with researchers using specific tools, or focusing on specific tasks within the hypothetico-deductive-inductive cycle of scientific inquiry. It is probably the work they specialize in, as opposed to fundamentally differing philosophical bents, that dictates which tradition a particular researcher will choose to follow. The following table summarizes the objectives encompassed by various types of research.

Research Tradition	Conceptual Approach	Objective	Task Control	Demands on Experiments and Data Acquisition	Examples of Methods
Interpretive / Qualitative	Synthetic / Holistic	Heuristic / Hypothesis Generating	Low, subjective and personal	Low	Specification Search, discovery, creation of new theory
Descriptive / Empirical	Analytical / Statistical	Hypothesis Testing / Theory Confirmation	Low, non-intrusive; deals with naturally occurring phenomena.	High	Confirmation of existing theory
Positive / Predictive	Analytic / Synthetic / Holistic	Accurate prediction and policy formulation	High if the goal is policy formulation and enactment	High	Schrödinger equation which predicts well despite controversy over what in nature it actually describes

This table looks at these objectives from a perspective more suitable to defining research disciplines. The character of available data, and its inherent observability often determine definability of model constructs. Important factors here are the manner in which we (1) explore available data in search of a model specification; (2) find out how well a model derived from existing theory is confirmed by a particular dataset; (3) discriminate one model from another by determining which one is better supported by the data; or (4) predict the existence and causal direction of potential relationships between candidate factors. Such analyses are respectively referred to as (1) specification search, or exploratory; (2) confirmatory; (3) discriminant and (4) causal-predictive or just plain predictive.

Type of Research	Objective	Typical Tasks	Typical Methods	Research Tradition
Exploratory	Explore available data in search of a model specification	Data reduction, pattern recognition	Human intuition and pattern recognition; statistical pattern recognition (neural nets, factor analysis)	Interpretive
Confirmatory	Find out how well a model derived from existing theory is confirmed by a particular dataset	Methods that measure how consistent observations are with theory	Statistical hypothesis testing	Descriptive-Empirical
Discriminant	Discriminate one model from another by determining which one is better supported by the data	Methods that measure how consistent observations are with one model versus another	Statistical hypothesis testing, pattern recognition, discriminant analysis	Descriptive-Empirical
Predictive	Predict the existence and causal direction of potential relationships between candidate factors.	Models that predict future observations, even if they seem not to be consistent with historical observations	Econometric forecasting models, neural networks	Positive

Though researchers may specialize, a full research program will likely incorporate interpretive, descriptive and positive phases. The appeal of latent variable SEM for studies in the social sciences is easy to understand. Many, if not most of the key concepts in the social sciences are not directly observable. The initial phases of nearly any research project typically involve a modicum of ad hoc pattern recognition. But at a certain point, structure is imposed, and the form that the research takes is determined by the: (1) Data that can be cost effectively acquired; (2) Hypotheses worth testing, and where they fit into larger theories; and (3) Objectives appropriate for the specific research at hand.
Tool selection, observation, survey instruments, data sources, statistics and reporting formats – are all dictated by these three factors.

Possibly the most compelling feature of modern path analysis tools such a PLS path analysis,

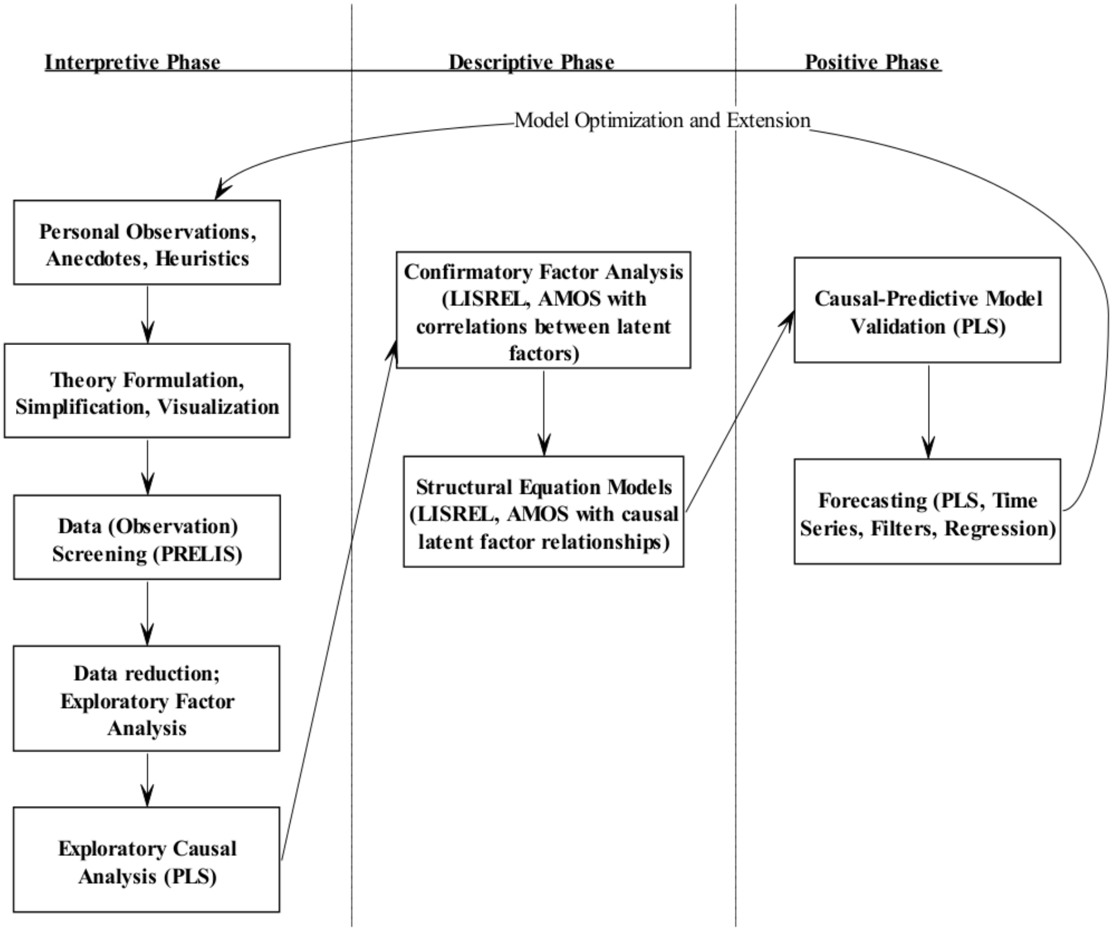
Interpretive Phase
Descriptive Phase
Positive Phase
Model Optimization and Extension
Personal Observations, Anecdotes, Heuristics
Theory Formulation, Simplification, Visualization
Data (Observation) Screening (PRELIS)
Data reduction; Exploratory Factor Analysis
Exploratory Causal Analysis (PLS)
Confirmatory Factor Analysis (LISREL, AMOS with correlations between latent factors)
Structural Equation Models (LISREL, AMOS with causal latent factor relationships)
Causal-Predictive Model Validation (PLS)
Forecasting (PLS, Time Series, Filters, Regression)

Figure 7.1: (#fig:rsch_pgm)Research Programs using SEM

AMOS and LISREL) is their ability to tease out network relationships of unobservable but theorized constructs. The choice of latent variable statistical methods arises through a dialectic arising from the need to cost effectively collect objective observations in research disciplines that mainly theorize about subjective and unobservable constructs. Without SEM path analysis tools, hypotheses testing tends to occur indirectly, leaving substantial opportunities for questioning their results and interpretation. Thus SEM path analysis approaches promise the direct testing of hypotheses about unobservables, but at a cost in complexity, and perhaps difficulty in interpreting exactly what the SEM statistical analysis concludes about each hypothesis.

Herman Wold suggested the concept of 'plausible causality,' a concept that was more completely developed in (Harris 1999). Wold's convergence on plausible causality took several turns over its development. Initially after Tukey's (Cochran, Mosteller, and Tukey 1954) suggestion that path analysis should adopt regression rather than correlation path coefficients, Wold explored both holistic analysis of variance approaches, along with piecewise path approaches. His main criticism of the analysis of variance approaches was that they filed to address non-Normal data, and that most data encountered by researchers is non-Normal, and often highly multicolinear (Hill 1979). Since you could not effectively render an opinion on whether the causal links in a model were true or not, one could at best conclude that these links were 'causally plausible.'

7.5 Hypothesis Testing

Hypotheses tests present a simplified model of the real world that can either be 'confirmed' or 'rejected' (thus the term 'confirmatory analyses') through analysis and summarization of data relevant to the underlying theory. No hypothesis can be unequivocally confirmed through data analysis; instead they are confirmed or rejected with some level of significance and power $1 - \beta$ (figure 3). The power of a statistical test is the probability that the test will reject a false null hypothesis. Significance and power $1 - \beta$, along with the distribution of the estimator will determine the minimum sample size that is required to perform the test.

Power analysis can either occur before (a priori) or after (ex post) data is collected. A priori power analysis is conducted prior to the conducting of research and is typically used to determine an appropriate sample size to achieve adequate power. Post-hoc power analysis is conducted after a study has been conducted and uses the obtained sample size and effect size to determine what the power was in the study assuming the effect size in the sample size is equal to the population effect size. Statistical power depends on:

1. The statistical significance criterion used in the test
2. The size of the difference or the strength of the similarity (that is, the effect size) in the population
3. The sensitivity of the data; or conversely, how much information the dataset has about the particular research question being studied in the hypothesis and theory.

Calculating the power requires first specifying the effect size you want to detect. The greater the effect size, the greater the power. Using statistical controls can increase sensitivity, by

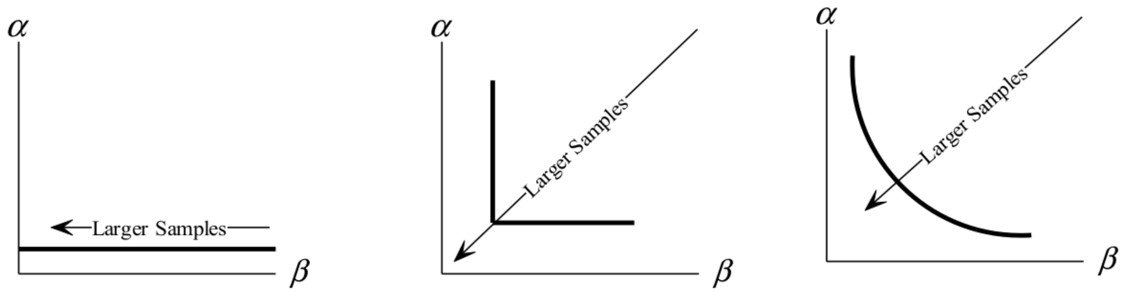

Figure 7.2: (#fig:rsch_trade)Choice and type I/II (i.e. error tradeoff with 1) fixed significance hypothesis tests; 2) minimax;3) cost-benefit (e.g., Bayes-risk) objective functions respectively from left to right in the figure

increasing the reliability of measures (as in psychometric reliability), and by increasing the size of the sample. Increasing sample size is the most commonly used method for increasing statistical power. Funding agencies, ethics boards and research review panels frequently request that a researcher perform a power analysis. The argument is that if a study is inadequately powered, there is no point in completing the research. Although there are no formal standards for power, most researchers who assess the power of their tests use 0.80 as a standard for adequacy.

7.6 Model Specification and Confirmation

Analysis of data will nearly always involve multiple steps of model specification, data collection and respecification. Model confirmation has two sides:

1. supporting the research hypotheses; and
2. rejecting competing theories. Difficulties in rejecting competing theories are one reason to favor simpler models (which lead to fewer competitor models).

If model richness and the inclusion of unobservable factors are required – which is the case in much of social science research – then analysis of variance and systems of regression methods improve over piecewise estimation in traditional and PLS path analysis because these imposes numerous restrictions on data collection (e.g., normality of distributions) and SEM definition (e.g., it is identified and theory driven). These restrictions mean that the researcher has to work harder for the estimates to be computed; but can make a much stronger case for model validity once the method is coaxed into generating an estimate. In many cases, the restrictions simply cannot be met – this is particularly true in the case of assuring that the observations are normally distributed. Censored, truncated, or ordinal data will not be normal (though one can test and argue for them being nearly normal). In such cases, weaker arguments may be generated or the researcher may just wait for either a better theory, or more data, or both. It is in addressing such combinatorial choice issues in descriptive research that the causal-positivist objectives of Wold's plausible causality are able to assist in model choice. Confirmatory tests of SEM models can be augmented by positive research that proposes alternative arrangements of latent factors, alternative causal relationships, and perhaps which adds new latent constructs while eliminating old. PLS path analysis is suited for such tasks because it does not demand knowledge of underlying distributions, equation identification, nor large datasets. It is designed to predict the strength of the model, and the strength of causal relationships between latent factors that can then be used to fine tune the hypotheses and the model tests for the next round of descriptive testing. PLS path analysis is cost-effective because of its very low demands on data – distributions need not be known, and estimates are generated even where there are many constructs in the model, and few observation, though this is due to bootstrapping by the computer algorithms, and is not the result of any inherent strength of the algorithms used for inference. Cost-effectiveness makes it a useful tool for prediction with available data, and for directing future descriptive research tasks such as data collection and model test setup (e.g., transformation of data to assure normality, inference of missing data, resolution of outliers and so forth) when current a priori knowledge for model building is weak, and where data acquisition opportunities are limited and expensive. As these are common challenges facing social scientists, we can easily see how PLS path analysis can fill an essential role in the path to building a convincing theory Wold (Noonan and Wold 1982) suggested that model 'confirmation' (i.e., the acceptance of a hypothesis as to its 'truth') occurs under a very broad range of circumstances, including small data sets and complex models, emphasizing that model predictions can only be considered plausible, rather than confirmed by the data testing. At the basis of this is the concept that rather than the researchers a priori hypothesized model being shown to be the only real model, we instead allow the existence of alternative models, indicating that this is one plausible model among several others. How many others? We look at this further in the next section.

7.7 How many alternative models should you test?

The stronger the correlations on an SEM path, the more power the method needs to have to detect an incorrect model. When correlations are low, the researcher may lack the power to reject the model at hand. Also, all competing SEM methods tend to overestimate goodness of fit for small samples of less that 200. Similarly, one can have good fit in a misspecified model. Equivalent models exist for almost all models, and the number of candidate models grows exponentially with the number of variables. Though systematic examination of equivalent models is still rare in practice, such examination is increasingly recommended. (Scheines et al. 1998) note, 'It is important to present all of the simplest alternatives compatible with the background knowledge and data rather than to arbitrarily choose one'. To gain a better insight into the alternative candidate models available, consider that confirmatory testing in descriptive research is basically a process of selecting one model over its alternatives. From a Neyman-Pearson perspective, it may be presented as a binary choice between accepting or rejecting a single (null) hypothesis; but in the larger research program, this has to be seen as choosing one hypothesized description of reality (i.e., the model) over many others.

How many? This depends on the complexity of hypotheses tested, and grows combinatorially large with the growth of alternative hypotheses. In the linked network models of SEM, as we showed in the prior chapter, the number of alternative hypotheses is proportional the square of the number of links.

7.8 Distributional Assumptions

Though the above studies provide some guidance on minimum sample sizes required for use of SEM in hypothesis testing, they do not address challenges of working with non-Normal data. This is a particular problem with SEM in the social science, because of the widespread use of Likert scale survey data 5 and 7 bin categorical data.

Jöreskog's maximum likelihood approach generates meaningful goodness-of-fit statistics for the latent SEM if the underlying indicator data is multinormal. PLS path analysis and Jöreskog's other algorithms don't explicitly impose this requirement, but the fit statistics are, unfortunately, difficult to interpret. We know from the studies cited previously that when PLS path analysis models are identified and the indicator data is multinormal, that parameter estimates converge to those of 2SLS, for which goodness-of-fit statistics are well understood. Indeed, (Goodhue, Lewis, and Thompson 2006) questioned whether the lack of robustness in the face of non-Gaussian observations should not be reason to abandon PLS path analysis for less lenient approaches such as LISREL.

(Noonan and Wold 1982) emphasized, though, that PLS-PA is robust, and will compute parameter estimates under a very broad range of circumstances, including small data sets and complex models. But beyond this, little is known of the small-sample properties of PLS-PA path estimators, or for that matter, how much these properties owe to bootstrapping in their computer implementations. Wold has promoted predictive -PA-PA use in the social sciences, because PLS-PA will yield predictions where other methods cannot, but these predictions can only be considered plausible, rather than confirmed by the data testing. If the data is both non-normal and has outliers, the decision to delete values or transform the data is confronted. Transformation of the variable is usually preferred as it typically reduces the number of outliers, and is more likely to produce normality, linearity, and homoscedasticity. In social science work, data is often limited to positive values only, and may be ordinal as well, as is the case for Likert scale responses. Screening will aid in the isolation of data peculiarities and allow the data to be adjusted in advance of further multivariate analysis (Tabachnick and Fidell, n.d.). When the researcher fails to Normalize data prior to analysis, then LISREL can search for the best latent variable model fit will error in several ways. If the maximum likelihood (ML) method is used, standard errors and statistics will be incorrect. In theory, weighted least squares procedures using the 'correct' weight matrix could produce correct estimates of standard errors and statistics, but will still require a substantially larger sample, and assurance that the weights have been properly chosen (which begets further questions).

7.9 Statistical Distributions: are they part of the Model or are they Part of the data?

Statistical analysis can be dichotomized into (1) analysis that assumes Normal or Gaussian distributions; and (2) so-called non-parametric analysis. This may oversimplify, but it does speak to the enormous temptation to 'assume' a Normal distribution, even where such assumptions are obviously violated.

Where other distributions arise, it is because the processes we are measuring, or the process of measurement itself, has a particular structure. Normal distributions tend to arise when we sum items. Other distributions arise in different situations , for example: (1) Binomial distributions are a result of binary – coin flipping – processes (2) Categorical distributions from polychotomous – dice throwing – processes; multinomial distributions from classification processes (3) Zipf-Pareto distributions from measuring rank-frequency (4) Lognormal distributions arise when we multiply items (5) Poisson distributions from Poisson processes

The point to bear in mind is that the distribution of the data depends to some extend on the way the experiment is conducted, the choice of factors to measure, and the processes that are associated with the reality we are investigating. Before blindly delving ahead and collecting data, it is important to see whether better behaved data (e.g., ones with a Normal distribution) can be acquired by altering the design of experiments or data collection.

7.10 Causality

The 18th century Scottish philosopher David Hume observed that 'causes are a strange sort of knowledge' (Hume 1758). People talk about causes and ostensibly the resulting effects as if they were facts about the structure of the universe, to be unraveled and used by scientists for the betterment of mankind. But causes aren't at all factual – rather they are in Hume's words a 'lively conception produced by habit.' They are the lazy conclusion, the scientific sound bite of modern life. Scientists can discover facts. But cause is not a fact. Cause is a simplification we use to make sense of the facts – it is scientific storytelling.

Causation is a mental shortcut that works well enough most of the time, and places minimal load on mental resources. Belief in causality allows us to safely drive a car (e.g., allowing us to predict that a crash will cause an injury) and otherwise make sense of the world. Unfortunately, when it comes to reasoning about complex, interrelated systems, where multiple correlations act simultaneously, cause is not fact. Cause is a mental shortcut that can be dangerously misleading.

In the 1940s, Flemish psychologist Albert Michotte explored the fabrication of causality in his experiments with observers of several short films about a red ball and a blue ball (Michotte and Thines 1991). In the first film, the red ball raced across the screen, touched the blue ball and then stopped. The blue ball, meanwhile, began moving in the same direction as the red ball. When asked to describe the film, viewers recalled that the red ball hit the blue ball, which caused it to move. Michotte called this the launching effect, and found it to be a universal property of visual perception. Although there was nothing about causation in the two-second film, viewers couldn't help but concoct a story about what had happened – they translated a sequence of stills into a causal story.

Michotte subsequently manipulated the films, asking subjects to describe the changes that took place in the new footage. For example, when he introduced a one-second break between the movement of the balls, the impression of causality disappeared. The red ball no longer 'caused' the blue ball to move. Michotte went on to conduct over one hundred similar experiments. For example in one, a small blue ball moved in front of a big red ball – subjects described this as the red ball 'chasing' the blue ball. If a big red ball was moving in front of a small blue ball, subjects described the blue ball as 'following' the red ball. Chasing and following were alternative story words that had replaced cause and effect. Michotte drew two conclusions from his experiments. First, our theories about cause and effect are inherently perceptual and subject to the visual shortcuts hardwired into our minds. Michotte saw causal beliefs as similar to color perception – particular objects will automatically be described as 'red' just as particular situations will be ascribed causality. Second, causality is a mental oversimplification, more useful in a day when humans regularly made split-second fight or flight decisions. The dangers of cause and effect simplifications have shown up most prominently – and expensively – in medical diagnosis and drug testing, particularly in the past decade. Pharmaceutical R&D is widely based on targeting causal links in metabolic pathways, and blocking them (an approach pioneered in the 19th century by Paul Erlich, which he called the 'magic bullet'). In contrast (Barabasi and Oltvai 2004) provide substantial evidence for perceiving these as metabolic networks rather than pathways, which revisits the arguments that lead from Sewall Wright's path analysis to more holistic approaches to SEM estimation. This matters today, because the R&D required discovering a new drug candidate is inflation adjusted one hundred times what it was in the 1950s. The average cost for approved drug molecule can from tens of millions to billions of dollars. Despite this, European regulators have found that 85% of approved new drugs work poorly or not at all. The consequences of oversimplification of metabolic pathways using chains of causal links has proved expensive in recent years. Lilly has in the past two years discarded two Alzheimer's drugs Semagacestat (which targets amyloid protein metabolism, but failed to do that, and also increased risk of skin cancer) and Dimebon (which targets mitochondria, but fails to slow Alzheimer's disease) after spending billions of dollars in R&D. Pfizer halted Phase III trials of cholesterol drug Torcetrapib, finding that it actually increased hear failure, with a 60% higher mortality rate.

How could these and many other studies have been so wrong? A fundamental failure to correctly model human metabolism. Metabolism is increasingly being understood to be a complex network of chemical interrelationships, where individual compounds may perform multiple service in multiple subsystems. Thinking of these in terms of linear metabolic pathways leads to incorrect models, and failed predictions of drug effectiveness. The simplifications of cause and effect result in misleading, wasteful and dangerous science that is costing the pharmaceutical industry dearly. Why is this important structural equation modeling? Because structural equation models almost invariably represent complex networks of inferred behavior that cannot be reduced to binary cause and effect relationships. If they could be reduced to simple relationships, we would estimate them as a series of regressions or ANOVA's. In modern society, the elaborate fictions of causality cost us dearly, and are a fundamental challenge in model building.

7.11 The Risks of Received Wisdom

Functional magnetic resonance imaging (fMRI) which measures change in blood flow related to neural activity in the brain and spinal cord has become a core tool in spinal diagnosis in the past two decades. Yet diagnosis has suffered a crisis of scientific method similar to that of drug R&D. Americans spend $90 billion annually treating back pain – roughly the same amount as is spent on cancer. Until the 1970s, the only remedy for back pain was bed rest, and most patients improved. With the advent of MRI in the 1990s, epidurals and surgery were increasingly prescribed, to treat the various 'causes' of back pain discovered in fMRI scans. Patient recovery declined.

Enormous amounts of data are generated by fMRI scans, and this data is error prone and difficult to interpret as brain processes are complex and often non-localized. Partial least squares structural models have been widely applied in fMRI imaging to analyze and interpret the mass of data generated in these 15-20 minute scans. Unfortunately, such analyses are only as accurate as the data, and are limited by the reliability of the brain model that used to describe the structural model's latent variables and factors.

The problem of fMRI data accuracy is illustrated in a widely cited article (Bennett, Wolford, and Miller 2009) that describes Dartmouth neuroscientist Craig Bennett's fMRI scan of a whole Atlantic salmon (purchased at a local fish market, and which, as Bennett dryly notes, 'was not alive at the time of scanning'). While the fish sat in the scanner, Bennett showed it 'a series of photographs depicting human individuals in social situations.' To maintain the rigor of the protocol the salmon, just like a human test subject, 'was asked to determine what emotion the individual in the photo must have been experiencing.' If that were all that had occurred, the salmon scanning would simply live on in Dartmouth lore as a 'crowning achievement in terms of ridiculous objects to scan.' But the fish had a surprise in store. When Bennett got around to analyzing the fMRI data, it looked as if the dead salmon was actually thinking about the pictures it had been shown. Not only has Bennett's study generated much mirth in an normally tedious field, but it has spawned its own slew of books – from both inside and outside of the field of brain imaging – critical of the statistical methods used to analyze all that data coming out of the fMRI machines (Bennett, Wolford, and Miller 2009). There are significant implications for use of fMRI in court, as fMRI scans have been promoted as alternatives to inadmissible polygraph evidence. One lesson should be taken away from these well-documented failures: all of the data collection and analysis in the world will not make up for the failings of a bad model. Social scientists are an insecure lot; and why wouldn't they be? Economics is dismissively referred to as 'the dismal science.' Physicists like Alan Sokal bait the community with faux articles such as 'Transgressing the Boundaries: Towards a Transformative Hermeneutics of Quantum Gravity' (Sokal 1996) (Sokal and Bricmont 1998). (Gross and Levitt 1997) in their book Higher Superstition have accused social scientists of practicing the black arts of post-modernist deconstructionism.

Thus it is unsurprising to find that abused social scientists sometimes suffer from physics envy – the preoccupation that every process natural or human, has a basis in something like Newtonian mechanics (despite the fact that modern physics has its quantum mechanics filled with vagaries). There is a tendency (perceived or real) for the so-called softer sciences and liberal arts to try to obtain mathematical expressions of their fundamental concepts, as an

attempt to move them closer to harder sciences, particularly physics. Yet the success of physics to mathematicize itself, particularly since Isaac Newton's Principia Mathematics, is generally considered as remarkable and often disproportionate compared to other areas of inquiry. Propensities towards complex graphs and unnecessary Greek notation are embarrassing symptoms of physics envy.
Science has traditionally bowed to the dictates of Occam's razor (lex parsimoniae) – the law of parsimony, economy or succinctness. It is a principle urging one to select among competing hypotheses that which makes the fewest assumptions and thereby offers the simplest explanation of the effect. Gratuitous complexity contributes to additional fallacies like:

1.Complex question (question presuppose the truth of some assumption buried in that question); 2. False cause (one treats as the cause of a thing what is not really the cause of that thing); 3. Apriorism (refusing to look at any evidence, such as plausible alternative models, that might count against one's claim or assumption); 4. Wishful thinking and 5. Composition (reasoning mistakenly from the attributes of a part to the attributes of the whole).

Nowhere are these fallacies more prominently on display than when an SEM path model becomes engorged with constructs – both latent and observed. Not just their sheer number of constructs distinguishes pretentious models, but also by their subjectivity. Tension, dissatisfaction, propensity, qualitative overload, pressure, scope, role conflict are all highly personal and highly subjective value judgments. Any illusion that an overburdened path analysis model is going to reveal deep insights is surely exaggerated. Modern software makes it easy to throw together complex models without any thought to their validity or usefulness – the computer can always figure out something. Still, the researcher would be well advised to be guided by Occam's Razor, or as Stephen Wolfram paraphrased 'it is vain to do with more what can be done with fewer' (Wolfram 2002b).

7.12 Design of Empirical Studies

7.12.1 Concepts

(Freedman 1987) objected to the SEM path analysis failure to distinguish among causal assumptions, statistical implications, and policy claims. His critique in a paper idiosyncratically titled "Statistical models and Shoe Leather" has ever since been a source of suspicion and confusion surrounding quantitative methods in the social sciences' captures to a great extent the general concerns of econometricians towards PLS path analysis and LISREL SEM statistical methods (as well as Wright's original correlation based path analysis). It arises from the concern that if the statistical methods alone don't insure well formed hypotheses, proper theory validation, and commensurate data analysis, that it is somehow flawed. But econometricians themselves live in the long tail, and are less constrained than other social sciences in their data analysis. The social sciences are at a disadvantage in data collection in comparison with the natural sciences. Historical records like financial statements and surveys of individual behavior tend not only to be subjective, but they are also one-shot, non-replicable measurements. Except in very contrived situations, social scientists find it difficult to set up a controlled lab experiment, and rerun it thousands or millions of times. Furthermore, central constructs in social science theory such as personal utilities are not directly observable. Purely statistical techniques can never solve these problems alone. Effort must be invested in arguing and formulating models and hypotheses as is dedicated to exploring alternative model specifications, predicting causes and consequences of a well-formed theory, and confirming theory with the data at hand. It is worth noting that even Bayesian statisticians are the target of variations on this objection. Because the Bayesians presume that there exists prior knowledge about the parameters being estimated (and the almost certainly is, if only the expected range of parameter values) and that knowledge can benefit the estimation. Somehow this raises suspicions that Bayesians will 'fudge' their priors to obtain a result. Whereas Bayesians explicitly separate out the subjective portion of their estimation, SEM methods and other econometric methods infuse that subjectivity into their hypothesis and underlying models.

SEM estimation place a heavy burden on theory formulation – the model contains both unobserved constructs as well as causal direction, with complex interactions between unobservables. SEM modeling is highly subjective, and thus the theories underlying SEM must be strong and well argued; alternatives must be proposed; and confirmatory testing needs to be extensive. Like the Bayesians perhaps, social scientists relying on SEM must adhere to a higher standard in their use of a priori subjective information. Subjective model and theory formulation needs to be subjected to a consequent greater scrutiny in the exploration and validation of that theory that one would find in the natural sciences or even in econometrics. In addition, many of the most interesting constructs in the social sciences are not directly observable. As a consequence, we are often forced to conjecture based on observations that we believe are correlated with these unobserved quantities – i.e., observed quantities that somehow 'indicate' what is going on with our unobserved 'latent' factors, thus they are called 'indicators.' The quest for knowledge in many of these important yet unobserved (latent) concepts (factors) usually takes one of three forms – either it is exploratory; theory confirmation; or predictive.

Interpretive and positive traditions involve the exploration of our observations for some simpler underlying set of factors. Descriptive theories represented in SEM hypothesize the latent factors underlying observations, and will conduct research to confirm or reject our hypotheses. Finally, we may only be interested in prediction – a much less demanding criterion than confirmation – and indeed may be willing to accept an incorrect model that nonetheless yield accurate predictions. The success of descriptive theory testing in the natural sciences – especially physics – in the 20th century have tended to bias our research expectations in the social sciences. Social sciences are disadvantaged by small datasets and inherently unobservable constructs underlying their most important theories. Such circumstances require an heavier investment in the inductive phases of research – interpretive and positive research – than in the theory confirmation of the descriptive phase. Current emphasis on the illusory rigor of descriptive research arises from a legacy of 'physics envy' dating back, one could argue, to Adolphe Quetelet in the early 19th century. But having developed the tools for more formal theory building, perhaps now is the time for the social sciences to reconsider their emphases, and devote more time to the induction of interpretive and positive research.

7.12.2 Significance testing

While many of the measures used in SEM can be assessed for significance, significance testing is less important in SEM than in other multivariate techniques. In other techniques, significance testing is usually conducted to establish that we can be confident that a finding is different from the null hypothesis, or, more broadly, that an effect can be viewed as 'real.' In SEM the purpose is usually to determine if one model conforms to the data better than an alternative model. It is acknowledged that establishing this does not confirm 'reality' as there is always the possibility that an unexamined model may conform to the data even better. More broadly, in SEM the focus is on the strength of conformity of the model with the data, which is a question of association, not significance. Other reasons why significance is of less importance in SEM:

1. SEM focuses on testing overall models, whereas significance tests are of single effects.
2. SEM requires relatively large samples. Therefore very weak effects may be found significant even for models which have very low conformity to the data.
3. SEM, in its more rigorous form, seeks to validate models with good fit by running them against additional (validation) datasets. Significance statistics are not useful as predictors of the likelihood of successful replication.

7.12.3 Model Identification

One way is to run a model-fitting program for pretest or fictional data, using your model. Model-fitting programs usually will generate error messages for underidentified models. As a rule of thumb, overidentified models will have degrees of freedom greater than zero in the chi-square goodness of fit test. AMOS has a tool icon to tell easily if degrees of freedom are positive. Some non-recursive models may also be identified (see (Kline 1998)). Degrees of freedom equal sample moments minus free parameters. The number of sample moments equals the number of variances plus covariances of indicator variables (for n indicator variables, this equals n(n+1)/2). The number of free parameters equals the sum of the number of error variances plus the number of factor (latent variable) variances plus the number of regression coefficients (not counting those constrained to be 1's).

7.12.4 Negative error variance estimates

When this occurs, your solution may be arbitrary. AMOS will give an error message saying that your solution is not admissible. LISREL will give an error message 'Warning: Theta EPS not positive definite.' Because the solution is arbitrary, modification indices, t-values, residuals, and other output cannot be computer or is arbitrary also. There are several reasons why one may get negative variance estimates.

1. This can occur as a result of high multicolinearity. Rule this out first.
2. Negative estimates may indicate Heywood cases (see below)

3. Even though the true value of the variance is positive, the variability in your data may be large enough to produce a negative estimate. The presence of outliers may be a cause of such variability. Having only one or two measurement variables per latent variable can also cause high standard errors of estimate.

For more on causes and handling of negative error variance, see (Chen et al. 2001)

7.12.5 Heywood cases

When the estimated error term for an indicator for a latent variable is negative, this nonsensical value is called a 'Heywood case.' Estimated variances of zero are also Heywood cases if the zero is the result of a constraint (without the constraint the variance would be negative). Heywood cases are typically caused by misspecification of the model, presence of outliers in the data, combining small sample size (ex., <100 or <150) with having only two indicators per latent variable, population correlations close to 1 or 0 (causing empirical underidentification), and/or bad starting values in maximum likelihood estimation. It is important that the final model not contain any Heywood cases. Solutions. Ordinarily the researcher will delete the offending indicator from the model, or will constrain the model by specifying a small positive value for that particular error term, and will otherwise work to specify a better-fitting model. Other strategies include dropping outliers from the data, applying nonlinear transforms to input data if nonlinear relations exist among variables, making sure there are at least three indicators per latent variable, specifying better starting values (better prior estimates), and gathering data on more cases. One may also drop MLE estimation in favor of GLS (generalized least squares) or even OLS (ordinary least squares).

7.12.6 Empirical Confirmation of Theory

In situations where theory is strong, confirmatory testing and model extension are goals, and appropriate datasets are available, the mainstream statistical approaches such as regression approaches developed at the Cowles Commission and Jöreskog's maximum likelihood SEM approaches are most appropriate. And for confirmatory testing, with well-understood fit indices and statistical measures, LISREL-ML is the tool of choice. But it extracts a high cost, a cost that is often prohibitive in investigating social science questions –underlying observations must be Multinormal; equation systems must be identified; and the model and data must actually converge to a solution. Very seldom are these conditions met even in the best of circumstances. Fortunately, Jöreskog has provided the PRELIS tools (and AMOS provides similar tools through its underlying SPSS) to filter and transform datasets so they meet these conditions, without robbing them of explanatory power. Nonetheless, for complex models of latent factors requires significantly more work and more expense than other methods discussed here.

	PLS Path Analysis	Covariance Structure Methods	Systems of Equations Regression
Ideal Applications	Prediction, specification search	Theory exploration and confirmation	Theory exploration and confirmation. hypothesis testing
Hypothesis testing?	n/a	Likelihood ratio test on observed vs. theoretical value of the dispersion matrix	Confidence interval procedures provides clearest roles for observations
Distributional assumption on indicators	None except that all indicators must have finite variance	Multinormal	Multinormal, but analysis of residuals and transformation allow options for non-Normal data
GUI	Yes	Yes	No
i.i.d. residuals?	No	Yes	Yes
Meaning of lines between latent factors	Canonical correlations	Covariances	Regression parameters on latent variables constructed from formative links
Full Information?	Limited	Full	Full
Solution process	Iterative search	Iterative Search	Closed form algebraic
Solution concept	Least squared error fit on pairs of variables	Maximum likelihood assuming Normal distribution	Least squared error fit
Fit measure and accuracy concept	No overall fit statistic		Many
Identifiabiltiy	No identification problem	Covariance structure is defined by the block structure of the model, the model may not be identified, and will have to be reparameterized	Rank condition, order condition

Chapter 8

Frontiers in Latent Variable Analysis

PLS-PA, LISREL and systems of regressions were designed for calculation on paper and with adding machines; they were disappointingly inadequate, but the best we had at the time. Statistical power has always lagged the size and complexity of the networks under analysis, and as a result generated unreliable, simplistic and inapplicable results. This is doubly unfortunate when we consider how important network models have reigned throughout mankind's history. For example: 1. The Romans were obsessed with water networks of aqueducts, plumbing and hydraulic networks. Romans visited public baths daily and upper class homes were centered on an interior pond. Hydraulic networks defined the medicine (bodily humors) science (hydraulics) and business systems (roads, canals, and pipes) of the Romans.
2. Hydraulic empires – Egypt, Somalia, the Ajuran Empire, Sri Lanka, Mesopotamia, China, Aztec, Maya and Indus Valley civilizations – were government structures of the largest ancient civilizations. These exercised power through exclusive control over access to water. Agricultural wealth, which accumulates around rivers and their arteries, created opportunities for hydraulic despotism through flood control and irrigation. Imperial bureaucracies required deep knowledge of hydraulic networks to rule and thrive (Wittfogel 1957) (Pryor 1982). 3. Hereditary and fealty networks defined the governments in the medieval world, and even into some 21st century regions. In feudal societies, politics and war were won or lost based on control of hereditary and fealty networks. 4. Genetic, metabolomic, proteomic, and epidemiological webs throughout the 20th century finally gave medicine a firm empirical and scientific foundation, allowing them to build on discoveries in chemistry, physics and mechanics.
5. In the 21st century, networks have so-far set the agenda for new paradigms in business, biology, sociology, computer science, finance, marketing and many other fields.

Each age has evolved its signature paradigms for linking the myriad networks structuring their worlds to the empirical reality deciding survival or extinction. The 21st century is differentiated by our acquisition of powerful network analysis tools using superfast computers with limitless storage directed by sophisticated algorithms. This chapter surveys the rapid evolution of computerized network analytics that hint at a deeper science that is only currently evolving.

8.1 Genetic Pathways Revisited.

Classical genetic mapping through pedigree analysis and breeding experiments could be used to determine sequence features within a genome, though the methods were time consuming with inherently low resolution. This was the world of Gregor Mendel and Sewall Wright; it was the world of 19th century dog breeders. In contrast, modern molecular gene mapping techniques are usually referred to as physical mapping – they use data from gene chips that measure which genes are active, or expressed, in a cell. Network analysis is providing a much more detailed and nuanced picture of life in all of its complexity. Amid thousands of studies using such chips, many compared the gene activity patterns in diseased tissue with that of healthy tissue. The number of genes associated with diseases is expanding rapidly because of so-called whole genome association studies. In these studies, gene chips are used to look for differences between the genomes of people with a disease and those without. Much of the raw data from such studies are deposited in databases, allowing researchers can now gather data on gene activity for scores of diseases and performs statistical analyses to map diseases based on similarities in their patterns of gene activity. Physicist Albert-László Barabási has been pushing the bounds of empirical analysis of networks for the past two decades, applying network theory to problems in the social sciences, commerce, physics, mathematics, and computer science. He has studied the growth and preferential attachment mechanisms responsible for the scale-free structure of the World Wide Web. But his most exciting studies have been in genetic networks, hearkening back to the origins of network path analysis by Gregor Mendel and Sewall Wright. Barabási and his colleagues obtained lists of disorders, disease genes, and their associations from the Online Mendelian Inheritance in Man database, compiling information on 1,286 disorders and 1,777 disease genes (Goh et al. 2007). Starting from a bipartite “diseasome” graph, they generated two network projections:

1. a human disease network that connected disorders to each other that share a common disease gene; and
2. a disease gene network that connected genes together that are associated with a common disorder. Diseases were represented by circles, or nodes, and linked to other diseases by lines that represent genes they have in common.

The human disease and gene network reveal the role of peripheral proteins for diseases caused by a variety of genetic mutations. Some diseases, such as Tay-Sachs, result from different mutations in a single gene, whereas other diseases, such as Zellweger syndrome, are caused by a mutation in any one of multiple genes. Generally, cancers were caused by somatic genetic mutations in essential or housekeeping genes. However, most inherited disease genes localized to the functional periphery of the network, with mutations preferentially in nonessential genes. Barabási’s research is changing the field of disease classification is known. Seemingly dissimilar diseases are being lumped together, and what were thought to be single diseases are being split into separate ailments. For example, two tumors that arise in the same part of the body and look the same on a pathologist’s slide might be quite different in terms of what is occurring at the gene and protein level. Certain breast cancers are already being treated differently from others because of genetic markers like estrogen receptor and Her2, and also more complicated patterns of genetic activity. Researchers can profiles drugs by the genes they activate as a way to find new uses for existing drugs. The research will also improve understanding of the causes of disease and of the functions of particular genes. For instance, two genes have recently been found to influence the risk of both diabetes and prostate cancer. But Barabási's network analysis advances medicine at a much more organic level – in providing a consistent way to classify diseases is also essential for tracking public health and detecting epidemics. The World Health Organization takes pains to periodically revise its International Classification of Diseases, which is used, among other ways, to tally the causes of death throughout the world. The classification is also the basis of the ICD-9 codes used for medical billing in the United States. The first international classification, in the 1850s, had about 140 categories of disease; the 10th edition, in 1993, had 12,000 categories. The increase stems mainly from better knowledge and diagnostic techniques that allow diseases to be distinguished from one another. For most of human history, diseases were named and classified by symptoms, which was all people could observe. Up to the eighteenth century, Aristotle and Galen were the primary references for medical knowledge. Linnaeus developed a symptom based taxonomy of disease with eleven classes — painful disease, motor diseases, blemishes and so on — that were further broken down into orders and species. Doctors who emphasized empirical observation, such as the surgeon John Hunter, were too often ignored by the medical establishment. By the nineteenth century surgery had advanced to the point where diseases began to be classified by their anatomic or physiological features. The stethoscope let doctors realize that what had been thought of as 17 conditions — like coughing up blood and shortness of breath — could all be different symptoms of the same disease, tuberculosis. Genetic networks allow the study of diseases at a finer level than even physiological tests. Genes are the instructions for the production of proteins, which interact in complex ways to carry out functions in the body. Disruptions in these molecular pathways can cause disease. Diseases have been subdivided by the type of mutation. Hemophilia was divided into hemophilia A and B, caused by mutations in different genes for different clotting factors. And what was once considered a mild form of hemophilia was later identified as a variant of a different clotting disorder, von Willebrand disease, caused by mutations in a different gene and requiring a different clotting factor as treatment. In contrast, two rare syndromes with different symptoms might represent a continuum of one disease. One syndrome, Meckel-Gruber, is tied to neural defects and death in babies. The other, Bardet-Biedl, is marked by vision loss, obesity, diabetes and extra fingers and toes.

8.2 Latent Constructs in Neural Networks

Neural networks have been subjects of much media speculation, where they are variously called machine-learning, artificial intelligence (AI) or deep-learning technologies. Their conflation with subjective human behaviors as well as over-promising and under-delivering are regular byproducts of all this attention, and it is often difficult to unwind signal from noise. Nonetheless, these evolving technologies offer great future promise in unraveling abstract, unobservable constructs that have been the mainstay of SEM research. Neural network applications in machine learning can be viewed as a natural extension, at least in their application, of the 20th century's generalized linear models articulated by John

Nelder (McCullagh and Nelder 1989) and Robert Wedderburn (Wedderburn 1974) combined with signal detection theory with signal processing filters (Schonhoff and Giordano 2006). Neural network models extend the 20th century matrix organization of data into tensors – multidimensional extensions of the traditional 2-dimensional matrices of observations and measurements. Whereas 20th century models model uncertainty in probability distributions with between one and four parameters that can be 'fitted' to data, 21st century neural network models allow models with a potentially unlimited set of parameters – the trainable weights in the model – that are optimized with a wide set of choices for optimization. The added dimensionality and increased number of trainable parameters requires much the larger datasets (what has been called 'big data' in the vernacular) that have been made possible by cloud services and several magnitudes increase in computing power obtained through parallelization. Nonetheless, problem conceptualizations parallel those of the traditional statistical setup, though the added flexibility of neural network models tends to favor model specification search and discovery over model confirmation in neural network based studies. Neural networks are built to discover rules to execute a data-processing task, given examples of what's expected. To do so requires four things: - Representations or measurements of the real world: these are transformations that represent or encode data from the real world into something that statistical or neural network models can work with. For instance, a color image can be encoded in several ways, for example a red-green-blue format or a hue-saturation-value format; both are representations of the same underlying real world data.
- Input data points: for example in image recognition or prediction, the input might be images.
- Examples of the expected output: for example in image recognition expected outputs might be tags such as a dog or a cat that are used to "train" the model. - Optimization criteria: optimization concepts generate the feedback signal that allows a neural network to "learn." It is not uncommon to hear neural networks criticized as black boxes: models that are difficult to extract and present in a human-readable form. In one sense this is true; but that is the price of power and complexity. Statistical models, for example SEM applications, that are complex or which deal extensively with unobservable constructs may be similarly difficult to interpret because of a lack of clear optimization or fit statistics. Nonetheless neural networks can provide powerful tools to extract latent constructs, even though we are still in the process of developing theory around their use.

A simple example, courtesy of demonstrates latent constructs in image recognition. Kaggle' public data platform presents competitions where anyone can create solutions to data science challenges. The "Dogs vs. Cats" competition is based on the Asirra (Animal Species Image Recognition for Restricting Access) dataset, collected by petfinder.com in collaboration with Microsoft. Such a challenge is often called a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) or HIP (Human Interactive Proof) and are used, among other things, to reduce email and blog spam and prevent brute-force attacks on web site passwords. Kaggle's training archive contains 25,000 pictures of dogs and cats. Currently, classification models score above 95% accuracy an the "Dogs vs. Cats" competition. The Kaggle competition in 2013 was won by entrants who used convolutional neural networks. The best entries achieved up to 95% accuracy.

The following is an example of one such convolutional network provided in (Chollet and Allaire 2018) that are amenable to visualization since they represent visual concepts to begin

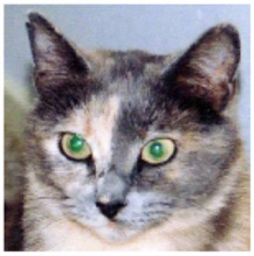

Figure 8.1: Cat image from Kaggle Dogs vs. Cats

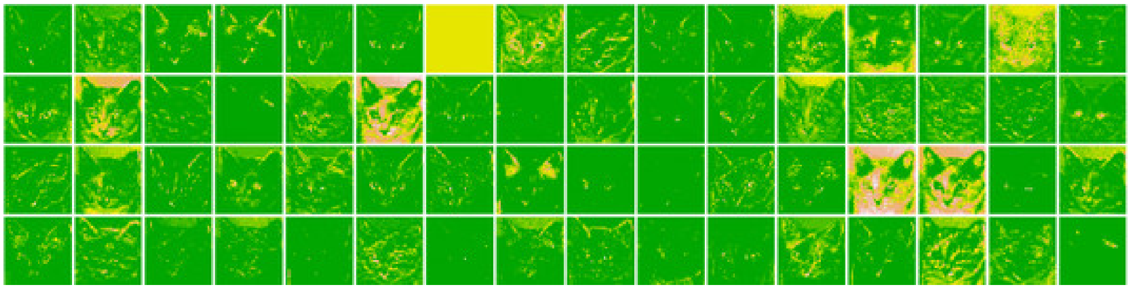

Figure 8.2: Cat image from Kaggle Dogs vs. Cats

with. The analysis starts by training the "Dogs vs. Cats" data on the @ref(fig.convnet) convolutional neural network model coded in Keras for R:

The original pictures are medium-resolution color JPEGs as in @ref(fig.cat)

The latent representations are represented in the channels in each of the layers of the neural network, as shown in the @ref(fig.catlat) for the input image @ref(fig.cat)

Most of these latent representations of cat features in @ref(fig.cat_latent) are difficult to interpret, though many of these seem to highlight features like edge and other topological characteristics that are unique to a cat image. Though the 'parameters' in a neural network must capture important features of a 'cat', since classification accuracy is over 95%, they are no doubt valid. But interpreting them in any human-readable form is difficult. Indeed, we may be moving into a statistical world where something besides fit statistics (such as R^2) is required to determine model validity.

8.3 The Evolution of SEM Research Questions

Sewall Wright's research was focused on teasing out an empirical understanding of Mendelian inheritance of biological traits long before there was any understanding of the basis underlying biological taxonomies. His objective was the creation of valid models of inheritance from raw data, without the benefit of preexisting theories. Observations could be replicated theoretically without end, simply by breeding another generation.
Applications of computer intensive canonical correlation to Wright's path analysis by Herman Wold and his student Karl Jöreskog were designed to prove or disprove theories about the

structural relationships between unobservable quantities in social sciences. Applications started with economics, but found greater usefulness in measuring unobservable model constructs such as intelligence, trust, value and so forth in psychology, sociology and consumer sentiment. Systems of regression equation approaches pioneered by Tjalling Koopmans were designed to prove or disprove theories about the structural relationships between economic measurements. Because regression fits actual observations rather than abstract concepts, they are able to provide a wealth of goodness-of-fit information to assess the quality of the theoretical models tested. Wold and Jöreskog's methods provide goodness-of-fit information that in comparison are sparse, unreliable, and difficult to interpret.

Social network analysis extends graph theory into empirical studies. It takes observations (e.g., Wright's genetic traits) and classifies them as 'nodes' (also called 'vertices'). It infers relationships, called 'edges' or 'links' between these nodes from real-world observations. Social networks reflect social relationships in terms of individuals (nodes) and relationships (links) between the individuals. Examples of links are contracts, acquaintances, kinship, employment and romantic relationships. A social network may be undirected, meaning that there is no distinction between the two nodes associated with each link, or its links may be directed from one node to another. In Wold's path analysis links are inherently undirected, but Wold invites researchers to 'plausibly infer' that links between unobserved variables have a direction (Basmann 1963a). Jöreskog's approach distinguishes between directed (path) links and undirected (confirmatory factor analysis) links between latent variables. Paths between observations and latent variables are always assumed to be directed. But in all of these cases, link direction derives from the researcher's a priori model building, not from empirical tests and data analysis. Research on social networks as a discipline began in the mid 19th century with the work of Karl Marx, Max Weber and David Émile Durkheim (Calhoun 2007). Mathematical approaches date from the mid-1950s with studies by Manfred Kochen, an Austrian who had been involved in urban design, and the interconnectedness and "social capital" of human networks, and is colleague Ithiel de Sola Pool a researcher on technology and society. Kochen and de Sola Pool's manuscript, Contacts and Influences (Sola Pool 1979) was conceived while both were working at the University of Paris in the early 1950s, during a time when psychologist Stanley Milgram visited and collaborated in their research. Michael Gurevich contributed empirical studies of the structure of social networks in his 1961 MIT dissertation under de Sola Pool which became part of their unpublished manuscript circulated among academics for over 20 years before publication in 1978. It formally articulated the mechanics of social networks, and explored the mathematical consequences of these, including the degree of connectedness. The manuscript left many significant questions about networks unresolved, and one of these was the number of degrees of separation in actual social networks. Milgram continued Gurevich's experiments in acquaintanceship networks at Harvard University on his return from Paris. His results were reported in "The Small World Problem" (Travers and Milgram 1967) in the popular science journal Psychology Today with a more rigorous version of the paper appearing in (Travers and Milgram 1977). The Psychology Today article generated enormous publicity for the experiments, which are well known today, long after much of the formative work has been forgotten. Milgram showed that people in the United States seemed to be connected by approximately three acquaintance links, on average.

Kochen and de Solla Poole subsequently constructed Monte Carlo simulations based on

Milgram's and Gurevich's data which recognized that both weak and strong acquaintance links are needed to model social structure. Kochen worked at IBM at the time, and the simulations, carried out on the relatively limited computers of the 1970s, were nonetheless able to predict that a more realistic three degrees of separation existed across the U.S. population. Their article "Contacts and Influences" concluded that in a U.S. sized population without social structure, "it is practically certain that any two individuals can contact one another by means of at most two intermediaries. In a [socially] structured population it is less likely but still seems probable. And perhaps for the whole world's population, probably only one more bridging individual should be needed." Their peers extrapolated these results to the well-known "six degrees of separation" for global population. As time progressed, Kochen and de Solla Poole's so-called 'small world' networks (Watts and Strogatz 1998) joined two other referent network structures: 'random' and 'scale-free' social network topologies (Barabási et al. 2003) which attempted to place empirical footings under a diverse set of social network topologies. Much of this work takes characteristics that have been studied in graph models, and matches them to particular empirical statistics from real-world networks. The are an important tool for research in sociology, political science, anthropology, biology, communications, finance, economics, bibliometrics, psychology, linguistics and marketing. The language of graph theory is rich with descriptors for network properties. Most of these can be applied to social networks such as those that have been the focus of this book. I will present here some of the most useful concepts germane to applications covered in this book. The interested researcher may follow up with a more extensive text of graph modeling to gain a more extensive understanding of the vocabulary of networks. Concepts important for path analysis and social network modeling fall into three categories:

1. visualization models that choose how to best present network information for human consumption;

2. link qualifiers and metrics that describe magnitude and qualitative features of relationships between nodes; and
3. topological statistics that summarize more fundamental geometric properties of the network.

8.4 Visualization: The New Language of Networks

Visualization for social networks has become popular with the development of software for mapping networks, particularly in 2-dimensions, for display. Dimensions may be added by dynamically adjusting parameter values on a time-lapse video, and by categorizing nodes and links with colors, shapes and legends. Visualization offers a powerful tool for publicizing data and research, but leaves open many opportunities for visual miscues, aberrations and illusions unchecked by more rigorous analytical procedures. Particular visualization approaches tend to be chosen their artistic appeal than for statistical veracity. Force-directed graph drawing algorithms provide one such popular approach for visualizing social networks. They position the nodes so that links are of equal length with few crossings, then assign spring-like forces to the links to place nodes at points of minimum ‘energy’ (Kobourov 2012). Fruchterman-Reingold (Fruchterman and Reingold 1991) Force Atlas 1 and 2 and other algorithms are used in software to produce attractive renderings of moderately large datasets. Larger datasets would great plots so dense with these methods that they would be unusable. In such cases, topological zooming algorithms render a level of detail dependent on the distance from one or more foci in order to achieve constant information density displays

References

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6). Ieee: 716–23.

Allen, I Elaine, and Christopher A Seaman. 2007. "Likert Scales and Data Analyses." *Quality Progress* 40 (7): 64–65.

Allen, I.E., and C.A. Seaman. 2007. "Likert Scales and Data Analyses." Journal Article. *Quality Progress* 40 (7): 64–65.

Altman, D.G., and P. Royston. 2000. "What Do We Mean by Validating a Prognostic Model?" Journal Article. *Statistics in Medicine* 19 (4): 453–73.

Anderson, J.C., and D.W. Gerbing. 1988. "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach." Journal Article. *Psychological Bulletin* 103 (3): 411.

Anderson, T.W., and H. Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." Journal Article. *The Annals of Mathematical Statistics* 20 (1): 46–63.

———. 1950. "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations." Journal Article. *The Annals of Mathematical Statistics*, 570–82.

Anderson, Theodore W, Herman Rubin, and others. 1950. "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics* 21 (4). Institute of Mathematical Statistics: 570–82.

Anderson, TW. 2005a. "Origins of the Limited Information Maximum Likelihood and Two-Stage Least Squares Estimators." Journal Article. *Journal of Econometrics* 127 (1): 1–16.

———. 2005b. "Origins of the Limited Information Maximum Likelihood and Two-Stage Least Squares Estimators." *Journal of Econometrics* 127 (1). Elsevier: 1–16.

Anderson, TW, N. Kunitomo, and Y. Matsushita. 2010. "On the Asymptotic Optimality of the Liml Estimator with Possibly Many Instruments." Journal Article. *Journal of*

Econometrics 157 (2): 191–204.

Bagozzi, Richard P, and Paul R Warshaw. 1990. "Trying to Consume." *Journal of Consumer Research* 17 (2). The University of Chicago Press: 127–40.

Barabasi, Albert-Laszlo, and Zoltan N Oltvai. 2004. "Network Biology: Understanding the Cell's Functional Organization." *Nature Reviews Genetics* 5 (2). Nature Publishing Group: 101.

Barabási, Albert-László, Zoltán Dezső, Erzsébet Ravasz, Soon-Hyung Yook, and Zoltán Oltvai. 2003. "Scale-Free and Hierarchical Structures in Complex Networks." In *AIP Conference Proceedings*, 661:1–16. 1. AIP.

Barclay, D., C. Higgins, and R. Thompson. 1995. "The Partial Least Squares (Pls) Approach to Causal Modeling: Personal Computer Adoption and Use as an Illustration." Journal Article. *Technology Studies* 2 (2): 285–309.

Basmann, Robert L. 1957. "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation." *Econometrica: Journal of the Econometric Society*. JSTOR, 77–83.

———. 1963a. "A Note on the Exact Finite Sample Frequency Functions of Generalized Classical Linear Estimators in a Leading Three-Equation Case." *Journal of the American Statistical Association* 58 (301). Taylor & Francis: 161–71.

———. 1963b. "The Causal Interpretation of Non-Triangular Systems of Economic Relations." *Econometrica: Journal of the Econometric Society*. JSTOR, 439–48.

Bastien, Philippe, Vincenzo Esposito Vinzi, and Michel Tenenhaus. 2005. "PLS Generalised Linear Regression." *Computational Statistics & Data Analysis* 48 (1). Elsevier: 17–46.

Benford, Frank. 1938. "The Law of Anomalous Numbers." *Proceedings of the American Philosophical Society*. JSTOR, 551–72.

Bennett, Craig M, George L Wolford, and Michael B Miller. 2009. "The Principled Control of False Positives in Neuroimaging." *Social Cognitive and Affective Neuroscience* 4 (4). Oxford University Press: 417–22.

Bentler, P.M. 1990. "Fit Indexes, Lagrange Multipliers, Constraint Changes and Incomplete Data in Structural Models." Journal Article. *Multivariate Behavioral Research* 25 (2): 163–72.

Bentler, P.M., and AB Mooijaart. 1989. "Choice of Structural Model via Parsimony: A Rationale Based on Precision." Journal Article. *Psychological Bulletin* 106 (2): 315.

Bentler, Peter M, and AB Mooijaart. 1989. "Choice of Structural Model via Parsimony: A Rationale Based on Precision." *Psychological Bulletin* 106 (2). American Psychological Association: 315.

Bielby, William Thomas, and Robert Mason Hauser. 1977. "Structural Equation Models." *Annual Review of Sociology* 3 (1). Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA: 137–61.

Bland, J Martin, and Douglas G Altman. 1995. "Multiple Significance Tests: The Bonferroni Method." Journal Article. *Bmj* 310 (6973): 170.

Blau, P.M., O.D. Duncan, and A. Tyree. 1967. "The Process of Stratification." Journal Article. *Social Stratification. Class, Race & Gender*, 317–29.

Blunch, N.J. 2008. *Introduction to Structural Equation Modelling Using Spss and Amos*. Book. Sage Publications Ltd.

Bollen, K.A. 1989. *Structural Equations with Latent Variables*. Book. Vol. 8. Wiley New York.

Bollen, Kenneth A. 1989. "A New Incremental Fit Index for General Structural Equation Models." *Sociological Methods & Research* 17 (3). Sage Publications: 303–16.

Bond, TG, and CM Fox. 2001. "Applying the Rasch Model. Fundamental Measurement in the Human Sciences.-Mahwah, New Jersy, Lawrence Erlbaum Associates, Inc."

Boomsma, A. 1982. "The Robustness of Lisrel Against Small Sample Sizes in Factor Analysis Models." Journal Article. *Systems Under Indirect Observation: Causality, Structure, Prediction* 1: 149–73.

———. 1985. "Nonconvergence, Improper Solutions, and Starting Values in Lisrel Maximum Likelihood Estimation." Journal Article. *Psychometrika* 50 (2): 229–42.

———. 1987. *The Robustness of Maximum Likelihood Estimation in Structural Equation Models*. Book. Cambridge University Press.

Box, George EP, and Norman R Draper. 2007. *Response Surfaces, Mixtures, and Ridge Analyses*. Vol. 649. John Wiley & Sons.

Browne, M.W., and R. Cudeck. 1989. "Single Sample Cross-Validation Indices for Covariance Structures." Journal Article. *Multivariate Behavioral Research* 24 (4): 445–55.

———. 1992. "Alternative Ways of Assessing Model Fit." Journal Article. *Sociological Methods & Research* 21 (2): 230–58.

———. 1993. "Alternative Ways of Assessing Model Fit." Journal Article. *SAGE FOCUS EDITIONS* 154: 136–36.

Browne, M.W., R. Cudeck, K. Tateneni, and G. Mels. 2002. "CEFA: Comprehensive Exploratory Factor Analysis." Journal Article. *Computer*.

Browne, Michael W, and Robert Cudeck. 1989. "Single Sample Cross-Validation Indices for Covariance Structures." *Multivariate Behavioral Research* 24 (4). Taylor & Francis: 445–55.

Burns, AC, and RF Bush. 2005. "Marketing Research: Online Research Applications. Person." prentice Hall.

Calhoun, Craig. 2007. *Nations Matter: Culture, History and the Cosmopolitan Dream*.

Routledge.

Cattell, Raymond B. 1966. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research* 1 (2). Taylor & Francis: 245–76.

Chan, Jason C. 1991. "Response-Order Effects in Likert-Type Scales." *Educational and Psychological Measurement* 51 (3). Sage Publications Sage CA: Thousand Oaks, CA: 531–40.

Chatelin, Yves-Marie, Vincenzo Esposito Vinzi, and Michel Tenenhaus. 2002. "State-of-Art on Pls Path Modeling Through the Available Software." Groupe HEC.

Chen, Feinian, Kenneth A Bollen, Pamela Paxton, Patrick J Curran, and James B Kirby. 2001. "Improper Solutions in Structural Equation Models: Causes, Consequences, and Strategies." *Sociological Methods & Research* 29 (4). Sage Publications: 468–508.

Cheung, Gordon W, and Roger B Rensvold. 2002. "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling* 9 (2). Taylor & Francis: 233–55.

Chin, W.W. 1998. "Commentary: Issues and Opinion on Structural Equation Modeling." Journal Article. *MIS Quarterly.*

———. 2010a. "Bootstrap Cross-Validation Indices for Pls Path Model Assessment." Journal Article. *Handbook of Partial Least Squares*, 83–97.

———. 2010b. "How to Write up and Report Pls Analyses." Journal Article. *Handbook of Partial Least Squares*, 655–90.

Chin, W.W., and J. Dibbern. 2010. "An Introduction to a Permutation Based Procedure for Multi-Group Pls Analysis: Results of Tests of Differences on Simulated Data and a Cross Cultural Analysis of the Sourcing of Information System Services Between Germany and the Usa." Journal Article. *Handbook of Partial Least Squares*, 171–93.

Chin, W.W., and P.R. Newsted. 1999. "Structural Equation Modeling Analysis with Small Samples Using Partial Least Squares." Journal Article. *Statistical Strategies for Small Sample Research* 2: 307–42.

Chin, WW, and J. Dibbern. 2009. "A Permutation Based Procedure for Multi-Group Pls Analysis: Results of Tests of Differences on Simulated Data and a Cross of Information System Services Between Germany and the Usa." Journal Article. *Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields, Berlin: Springer.*

Chollet, François, and Joseph J Allaire. 2018. *Deep Learning with R, Ch.5.4.* Manning Publications Company.

Christ, Carl F. 1994. "The Cowles Commission's Contributions to Econometrics at Chicago, 1939-1955." *Journal of Economic Literature* 32 (1). JSTOR: 30–59.

Clarke, Shelley, Jonathan Worcester, Glen Dunlap, Marcey Murray, and Kathy Bradley-Klug. 2002. "Using Multiple Measures to Evaluate Positive Behavior Support: A Case Example."

Journal of Positive Behavior Interventions 4 (3). Sage Publications Sage CA: Los Angeles, CA: 131–45.

Cochran, William G, Frederick Mosteller, and John W Tukey. 1954. “Principles of Sampling.” *Journal of the American Statistical Association* 49 (265). Taylor & Francis: 13–35.

Cohen, Jacob. 1988. “Statistical Power Analysis for the Behavioral Sciences 2nd Edn.” Erlbaum Associates, Hillsdale.

Copas, John B, and HG Li. 1997. “Inference for Non-Random Samples.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59 (1). Wiley Online Library: 55–95.

Cowles 3rd, Alfred, and Edward N Chapman. 1935. “A Statistical Study of Climate in Relation to Pulmonary Tuberculosis.” *Journal of the American Statistical Association* 30 (191A). Taylor & Francis: 517–36.

Cowles, A. 1933. “Can Stock Market Forecasters Forecast?” Journal Article. *Econometrica: Journal of the Econometric Society*, 309–24.

Cox III, Eli P. 1980. “The Optimal Number of Response Alternatives for a Scale: A Review.” *Journal of Marketing Research.* JSTOR, 407–22.

Critchley, Hugo D. 2002. “Electrodermal Responses: What Happens in the Brain.” *The Neuroscientist* 8 (2). SAGE Publications Sage CA: Los Angeles, CA: 132–42.

Critchley, Hugo D, Raphael N Melmed, Eric Featherstone, Christopher J Mathias, and Raymond J Dolan. 2002. “Volitional Control of Autonomic Arousal: A Functional Magnetic Resonance Study.” *Neuroimage* 16 (4). Elsevier: 909–19.

Davis, Fred D, Richard P Bagozzi, and Paul R Warshaw. 1989. “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models.” *Management Science* 35 (8). INFORMS: 982–1003.

———. 1992. “Extrinsic and Intrinsic Motivation to Use Computers in the Workplace 1.” *Journal of Applied Social Psychology* 22 (14). Wiley Online Library: 1111–32.

Dawes, Sharon S, and Anthony M Cresswell. 2012. “From ‘Need to Know’ to ‘Need to Share’: Tangled Problems, Information Boundaries, and the Building of Public Sector Knowledge Networks.” In *Debating Public Administration*, 93–114. Routledge.

Dhrymes, Phoebus J. 1972. “Distributed Lags: A Survey.” Report. UCLA Department of Economics.

———. 1974. *Econometrics.* Book. Springer.

Dhrymes, Phoebus J, and H Erlat. 1972. “Asymptotic Properties of Full Information Estimators in Dynamic Autoregressive Simultaneous Models.” Report. UCLA Department of Economics.

Dhrymes, Phoebus J, E Philip Howrey, Saul H Hymans, Jan Kmenta, Edward E Leamer, Richard E Quandt, James B Ramsey, Harold T Shapiro, and Victor Zarnowitz. 1972. “Criteria for Evaluation of Econometric Models." Book Section. In *Annals of Economic and Social Measurement, Volume 1, Number 3*, 291–325. NBER.

Dietz, Laura, Steffen Bickel, and Tobias Scheffer. 2007. "Unsupervised Prediction of Citation Influences." In *Proceedings of the 24th International Conference on Machine Learning*, 233–40. ACM.

Dijkstra, Theo. 1983. "Some Comments on Maximum Likelihood and Partial Least Squares Methods." *Journal of Econometrics* 22 (1-2). Elsevier: 67–90.

Ding, L., W.F. Velicer, and L.L. Harlow. 1995. "Effects of Estimation Methods, Number of Indicators Per Factor, and Improper Solutions on Structural Equation Modeling Fit Indices." Journal Article. *Structural Equation Modeling: A Multidisciplinary Journal* 2 (2): 119–43.

Farebrother, R.W. 1999. *Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900*. Book. Springer Verlag.

Féré, Charles. 1899. *The Pathology of Emotions*. University Press.

Fisher, Ronald A. 1921. "On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample." *Metron* 1: 3–32.

Fisher, Ronald Aylmer. 2006. *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd (original 1925).

Fornell, Claes, and D Larker. 1981. "Structural Equation Modeling and Regression: Guidelines for Research Practice." *Journal of Marketing Research* 18 (1): 39–50.

Fowles, Don C, Margaret J Christie, Robert Edelberg, William W Grings, David T Lykken, and Peter H Venables. 1981. "Publication Recommendations for Electrodermal Measurements." *Psychophysiology* 18 (3). Wiley Online Library: 232–39.

Fox, J. 2006. "Structural Equation Modeling with the Sem Package in R." Journal Article. *Structural Equation Modeling* 13 (3): 465–86.

Fox, John. 2002. "Structural Equation Models." *Appendix to an R and S-PLUS Companion to Applied Regression.*

Frank, I.E., and J.H. Friedman. 1993. "A Statistical View of Some Chemometrics Regression Tools." Journal Article. *Technometrics*, 109–35.

Freedman, David A. 1987. "As Others See Us: A Case Study in Path Analysis." *Journal of Educational Statistics* 12 (2). SAGE Publications Sage CA: Los Angeles, CA: 101–28.

Freedman, David H. 2010. "Lies, Damned Lies, and Medical Science." *The Atlantic* 306 (4): 76–84.

Friedman, Hershey H, and Taiwo Amoo. 1999. "Rating the Rating Scales."

Friedman, Hershey H, Yonah Wilamowsky, and Linda W Friedman. 1981. "A Comparison of Balanced and Unbalanced Rating Scales." *The Mid-Atlantic Journal of Business* 19 (2): 1–7.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for

Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1). NIH Public Access: 1.

Friedman, Jerome, Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. 2004. "Discussion of Boosting Papers." *Ann. Statist* 32: 102–7.

Fruchterman, Thomas MJ, and Edward M Reingold. 1991. "Graph Drawing by Force-Directed Placement." *Software: Practice and Experience* 21 (11). Wiley Online Library: 1129–64.

Gerbing, D.W., and J.C. Anderson. 1988. "An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment." Journal Article. *Journal of Marketing Research*, 186–92.

Geweke, J.F., and K.J. Singleton. 1981. "Maximum Likelihood' Confirmatory' Factor Analysis of Economic Time Series." Journal Article. *International Economic Review* 22 (1): 37–54.

Gilbert, Paul. 1998. "The Evolved Basis and Adaptive Functions of Cognitive Distortions." Journal Article. *British Journal of Medical Psychology* 71 (4): 447–63.

Goh, Kwang-Il, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. 2007. "The Human Disease Network." *Proceedings of the National Academy of Sciences* 104 (21). National Acad Sciences: 8685–90.

Goodhue, D., W. Lewis, and R. Thompson. n.d. "PLS, Small Sample Size, and Statistical Power in Mis Research." Conference Proceedings. In, 8:202b–202b. IEEE.

Goodhue, Dale L. 1995. "Understanding User Evaluations of Information Systems." *Management Science* 41 (12). INFORMS: 1827–44.

Goodhue, Dale L, and Ronald L Thompson. 1995. "Task-Technology Fit and Individual Performance." *MIS Quarterly*. JSTOR, 213–36.

Goodhue, Dale L, William Lewis, and Ron Thompson. 2012a. "Comparing Pls to Regression and Lisrel: A Response to Marcoulides, Chin, and Saunders." *Mis Quarterly*. JSTOR, 703–16.

———. 2012b. "Does Pls Have Advantages for Small Sample Size or Non-Normal Data?" *Mis Quarterly*. JSTOR, 981–1001.

Goodhue, Dale, William Lewis, and Ron Thompson. 2006. "PLS, Small Sample Size, and Statistical Power in Mis Research." In *System Sciences, 2006. Hicss'06. Proceedings of the 39th Annual Hawaii International Conference on*, 8:202b–202b. IEEE.

Gould, Stephen Jay. 1977. *Ontogeny and Phylogeny*. Harvard University Press.

Greco, Alberto, Gaetano Valenza, and Enzo Pasquale Scilingo. 2016. *Advances in Electrodermal Activity Processing with Applications for Mental Health*. Springer.

Greco, Alberto, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2016. "CvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing."

IEEE Transactions on Biomedical Engineering 63 (4). IEEE: 797–804.

Grier, David Alan. 2013. *When Computers Were Human.* Princeton University Press.

Gross, Paul R, and Norman Levitt. 1997. *Higher Superstition: The Academic Left and Its Quarrels with Science.* JHU Press.

Hair, Joe F, Christian M Ringle, and Marko Sarstedt. 2011. "PLS-Sem: Indeed a Silver Bullet." *Journal of Marketing Theory and Practice* 19 (2). Taylor & Francis: 139–52.

Hair, Joe F, Marko Sarstedt, Christian M Ringle, and Jeannette A Mena. 2012. "An Assessment of the Use of Partial Least Squares Structural Equation Modeling in Marketing Research." Journal Article. *Journal of the Academy of Marketing Science* 40 (3): 414–33.

Harris, R.L. 1999. *Information Graphics: A Comprehensive Illustrated Reference.* Book. Oxford University Press. http://books.google.com/books?id=LT1RXREvkGIC.

Hauser, Robert M. 1972. "Disaggregating a Social-Psychological Model of Educational Attainment." *Social Science Research* 1 (2). Elsevier: 159–88.

Hauser, Robert M, and Arthur S Goldberger. 1971. "The Treatment of Unobservable Variables in Path Analysis." *Sociological Methodology* 3. JSTOR: 81–117.

He, Qinying, and HN Nagaraja. 2011. "Correlation Estimation in Downton's Bivariate Exponential Distribution Using Incomplete Samples." *Journal of Statistical Computation and Simulation* 81 (5). Taylor & Francis: 531–46.

Headon, Robert, and Rupert Curwen. 2002. "Movement Awareness for Ubiquitous Game Control." *Personal and Ubiquitous Computing* 6 (5-6). Springer: 407–15.

Henrich, Joseph, and Richard McElreath. 2003. "The Evolution of Cultural Evolution." Journal Article. *Evolutionary Anthropology: Issues, News, and Reviews* 12 (3): 123–35.

Henseler, J., and G. Fassott. 2010. "Testing Moderating Effects in Pls Path Models: An Illustration of Available Procedures." Journal Article. *Handbook of Partial Least Squares*, 713–35.

Henseler, J., C.M. Ringle, and R.R. Sinkovics. 2009. "The Use of Partial Least Squares Path Modeling in International Marketing." Journal Article. *Advances in International Marketing* 20 (2009): 277–319.

Hill, Bruce M. 1979. "Posterior Moments of the Number of Species in a Finite Population and the Posterior Probability of Finding a New Species." Journal Article. *Journal of the American Statistical Association* 74 (367): 668–73.

———. 1992. "Bayesian Nonparametric Prediction and Statistical Inference." Book Section. In *Bayesian Analysis in Statistics and Econometrics*, 43–94. Springer.

Hill, Theodore P. 1995. "Base-Invariance Implies Benford's Law." *Proceedings of the American Mathematical Society* 123 (3): 887–95.

Hotelling, H. 1936. "Relations Between Two Sets of Variates." Journal Article. *Biometrika*

28 (3/4): 321–77.

Hotelling, Harold. 1933. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology* 24 (6). Warwick & York: 417.

Hox, Joop J. 1995. *Applied Multilevel Analysis.* TT-publikaties.

Hsu, Chien-Lung, J Christopher Westland, and Chun-Hao Chiang. 2015. "Electronic Commerce Research in Seven Maps." *Electronic Commerce Research* 15 (2). Springer: 147–58.

Hu, L., and P.M. Bentler. 1999. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." Journal Article. *Structural Equation Modeling: A Multidisciplinary Journal* 6 (1): 1–55.

Hume, David. 1758. *Essays and Treatises on Several Subjects.* A. Millar;; A. Kincaid; A. Donaldson, at Edinburgh.

Ioannidis, J.P.A. 2005. "Why Most Published Research Findings Are False." Journal Article. *PLoS Medicine* 2 (8): e124.

Ioannidis, JP. 2005. "Differentiating Biases from Genuine Heterogeneity: Distinguishing Artifactual from Substantive Effects." *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments.* Wiley Chichester, UK, 287–302.

James, William. 1884. "What Is an Emotion?" *Mind* 9 (34). JSTOR: 188–205.

Jamieson, Susan. 2004. "Likert Scales: How to (Ab) Use Them." *Medical Education* 38 (12). Wiley Online Library: 1217–8.

Joreskog, K.G., and M. Van Thillo. 1972. "LISREL: A General Computer Program for Estimating a Linear Structural Equation System Involving Multiple Indicators of Unmeasured Variables." Journal Article.

Joreskog, Karl G. 1970. "A General Method for Estimating a Linear Structural Equation System." Journal Article.

Joreskog, Karl G, Dag Sorbom, and Jay Magidson. 1979. "Advances in Factor Analysis and Structural Equation Models." Journal Article.

Jöreskog, Karl G. 1993. "Testing Structural Equation Models." Journal Article. *Sage Focus Editions* 154: 294–94.

Jöreskog, Karl G, and Arthur S Goldberger. 1975. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." *Journal of the American Statistical Association* 70 (351a). Taylor & Francis: 631–39.

Jöreskog, Karl G, and Dag Sörbom. 1982. "Recent Developments in Structural Equation Modeling." Journal Article. *Journal of Marketing Research*, 404–16.

Jung, Carl Gustav. 1919. *Studies in Word-Association: Experiments in the Diagnosis of Psychopathological Conditions Carried Out at the Psychiatric Clinic of the University of*

Zurich. Moffat, Yard & Co.

Kahai, S.S., and R.B. Cooper. 2003. “Exploring the Core Concepts of Media Richness Theory: The Impact of Cue Multiplicity and Feedback Immediacy on Decision Quality.” Journal Article. *Journal of Management Information Systems* 20 (1): 263–99.

Kailath, Thomas. 1967. “The Divergence and Bhattacharyya Distance Measures in Signal Selection.” *IEEE Transactions on Communication Technology* 15 (1). IEEE: 52–60.

Kaiser, Henry F. 1960. “The Application of Electronic Computers to Factor Analysis.” *Educational and Psychological Measurement* 20 (1). Sage Publications Sage CA: Thousand Oaks, CA: 141–51.

Kendall, Maurice, and JDR Gibbons. 1990. “Correlation Methods.” Oxford: Oxford University Press.

Kline, Rex B. 1998. “Software Review: Software Programs for Structural Equation Modeling: Amos, Eqs, and Lisrel.” *Journal of Psychoeducational Assessment* 16 (4). Sage Publications Sage CA: Thousand Oaks, CA: 343–64.

Kobourov, Stephen G. 2012. “Spring Embedders and Force Directed Graph Drawing Algorithms.” *arXiv Preprint arXiv:1201.3011.*

Komorita, Samuel S, and William K Graham. 1965. “Number of Scale Points and the Reliability of Scales.” *Educational and Psychological Measurement* 25 (4). Sage Publications Sage CA: Thousand Oaks, CA: 987–95.

Koopmans, Tjalling C. 1951. “Analysis of Production as an Efficient Combination of Activities.” Journal Article. *Activity Analysis of Production and Allocation* 13: 33–37.

Koopmans, Tjalling C, and Martin Beckmann. 1957. “Assignment Problems and the Location of Economic Activities.” *Econometrica: Journal of the Econometric Society.* JSTOR, 53–76.

Kühberger, Anton. 1995. “The Framing of Decisions: A New Look at Old Problems.” *Organizational Behavior and Human Decision Processes* 62 (2). Elsevier: 230–40.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Vol. 53. John Wiley & Sons Incorporated.

Lee, Jerry W, Patricia S Jones, Yoshimitsu Mineyama, and Xinwei Esther Zhang. 2002. “Cultural Differences in Responses to a Likert Scale.” *Research in Nursing & Health* 25 (4). Wiley Online Library: 295–306.

Liberati, Alessandro, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Pl J Devereaux, Jos Kleijnen, and David Moher. 2009. “The Prisma Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration.” *PLoS Medicine* 6 (7).

Public Library of Science: e1000100.

Likert, Rensis. 1932. “A Technique for the Measurement of Attitudes.” *Archives of Psychology.*

———. 1961. “New Patterns of Management.” McGraw-Hill.

Lohmoller, Jan-Bernd. 1988. “The Pls Program System: Latent Variables Path Analysis with Partial Least Squares Estimation.” Journal Article. *Multivariate Behavioral Research* 23 (1): 125–27.

Lohmöller, Jan-Bernd. 1989. *Latent Variable Path Modeling with Partial Least Squares.* Book. Physica-Verlag Heidelberg.

Lucas Jr, Robert E. 1992. “Econometric Policy Evaluation: A Critique.” *The New Classical Macroeconomics*, 3–30.

Ludden, Thomas M, Stuart L Beal, and Lewis B Sheiner. 1994. “Comparison of the Akaike Information Criterion, the Schwarz Criterion and the F Test as Guides to Model Selection.” *Journal of Pharmacokinetics and Biopharmaceutics* 22 (5). Springer: 431–45.

Lydtin, H, G Lohmöller, R Lohmöller, H Schmitz, and I Walter. 1980. “Hemodynamic Studies on Adalat in Healthy Volunteers and in Patients.” Conference Proceedings. In *2nd International Adalat® Symposium*, 112–23. Springer.

Mandelbrot, Benoit B. 1982. *The Fractal Geometry of Nature.* Vol. 1. WH freeman New York.

Marcoulides, George A, and Carol Saunders. 2006. “Editor’s Comments: PLS: A Silver Bullet?” *MIS Quarterly.* JSTOR, iii–ix.

Marsh, H.W., and M. Bailey. 1991. “Confirmatory Factor Analyses of Multitrait-Multimethod Data: A Comparison of Alternative Models.” Journal Article. *Applied Psychological Measurement* 15 (1): 47.

Marsh, H.W., and A.S. Yeung. 1997. “Causal Effects of Academic Self-Concept on Academic Achievement: Structural Equation Models of Longitudinal Data.” Journal Article. *Journal of Educational Psychology* 89 (1): 41.

———. 1998. “Longitudinal Structural Equation Models of Academic Self-Concept and Achievement: Gender Differences in the Development of Math and English Constructs.” Journal Article. *American Educational Research Journal* 35 (4): 705.

Marsh, H.W., B.M. Byrne, and R. Craven. 1992. “Overcoming Problems in Confirmatory Factor Analyses of Mtmm Data: The Correlated Uniqueness Model and Factorial Invariance.” Journal Article. *Multivariate Behavioral Research* 27 (4): 489–507.

Marsh, H.W., Z. Wen, and K.T. Hau. 2004. “Structural Equation Models of Latent Interactions: Evaluation of Alternative Estimation Strategies and Indicator Construction.” Journal Article. *Psychological Methods* 9 (3): 275.

Marsh, Herbert W, and Michael Bailey. 1991. “Confirmatory Factor Analyses of Multitrait-Multimethod Data: A Comparison of Alternative Models.” *Applied Psychological Measurement*

15 (1). Sage Publications Sage CA: Thousand Oaks, CA: 47–70.

Marston, William Moulton. 1938. “The Lie Detector Test.” RR Smith.

Matell, Michael S, and Jacob Jacoby. 1972. “Is There an Optimal Number of Alternatives for Likert-Scale Items? Effects of Testing Time and Scale Properties.” *Journal of Applied Psychology* 56 (6). American Psychological Association: 506.

Matsumoto, Makoto, and Takuji Nishimura. 1998. “Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator.” *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 8 (1). ACM: 3–30.

McArdle, Brian H, and Marti J Anderson. 2001. “Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis.” *Ecology* 82 (1). Wiley Online Library: 290–97.

McArdle, John J. 1988. “Dynamic but Structural Equation Modeling of Repeated Measures Data.” Book Section. In *Handbook of Multivariate Experimental Psychology*, 561–614. Springer.

McArdle, John J, and David Epstein. 1987. “Latent Growth Curves Within Developmental Structural Equation Models.” Journal Article. *Child Development*, 110–33.

McCarthy, Mark I, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. 2008. “Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges.” *Nature Reviews Genetics* 9 (5). Nature Publishing Group: 356.

McCullagh, Peter, and John A Nelder. 1989. *Generalized Linear Models*. Vol. 37. CRC press.

Merton, R.K. 1968. “The Matthew Effect in Science.” Journal Article. *Science* 159 (3810): 56.

Merton, Robert K. 1988. “The Matthew Effect in Science, Ii: Cumulative Advantage and the Symbolism of Intellectual Property.” *Isis* 79 (4). Department of History; Science, University of Pennsylvania: 606–23.

———. 1995. “The Thomas Theorem and the Matthews Effect.” *Soc. F.* 74. HeinOnline: 379.

Merton, Robert King, and Robert C Merton. 1968. *Social Theory and Social Structure*. Simon; Schuster.

Meshkati, Najmedin, Peter A Hancock, Mansour Rahimi, and Suzanne M Dawes. 1995. “Techniques in Mental Workload Assessment.” Taylor & Francis.

Michotte, Albert, and Georges Thines. 1991. “Perceived Causality.” *Michotte’s Experimental Phenomenology of Perception*. Lawrence Erlbaum Hove, East Sussex, 66–87.

Monecke, Armin, and Friedrich Leisch. 2012a. “SemPLS: Structural Equation Modeling Using Partial Least Squares.” Journal Article. *Journal of Statistical Software* 48 (3): 1–32.

———. 2012b. “semPLS: Structural Equation Modeling Using Partial Least Squares.” *Journal* *of Statistical Software* 48 (3): 1–32. http://www.jstatsoft.org/v48/i03/.

Murphy, Gardner, and Rensis Likert. 1938. *Public Opinion and the Individual.* Harper.

Nakagawa, Shinichi, and Innes C Cuthill. 2007. "Effect Size, Confidence Interval and Statistical Significance: A Practical Guide for Biologists." *Biological Reviews* 82 (4). Wiley Online Library: 591–605.

Neuhoff, John G. 2001. "An Adaptive Bias in the Perception of Looming Auditory Motion." Journal Article. *Ecological Psychology* 13 (2): 87–110.

Neumann, Eva, and Richard Blanton. 1970. "The Early History of Electrodermal Research." *Psychophysiology* 6 (4). Wiley Online Library: 453–75.

Noonan, Richard, and Herman Wold. 1982. "PLS Path Modeling with Indirectly Observed Variables: A Comparison of Alternative Estimates for the Latent Variable." *Systems Under Indirect Observation, Part II.* Amsterdam: North Holland.

Norman, Geoff. 2010. "Likert Scales, Levels of Measurement and the 'Laws' of Statistics." *Advances in Health Sciences Education* 15 (5). Springer: 625–32.

Nunnally, J.C. 1967. *Psychometric Theory.* Book. Tata McGraw-Hill Education.

Nunnally, Jum C, Ira H Bernstein, and others. 1967. *Psychometric Theory.* Vol. 226. McGraw-Hill New York.

Picard, Rosalind W. 2010. "Affective Computing: From Laughter to Ieee." *IEEE Transactions on Affective Computing* 1 (1). IEEE: 11–17.

Podsakoff, Philip M, and Dennis W Organ. 1986. "Self-Reports in Organizational Research: Problems and Prospects." *Journal of Management* 12 (4). Sage Publications Sage CA: Thousand Oaks, CA: 531–44.

Pryor, Robert. 1982. "Values, Preferences, Needs, Work Ethics, and Orientations to Work: Toward a Conceptual and Empirical Integration." *Journal of Vocational Behavior* 20 (1). Elsevier: 40–52.

Ramaratnam, Sridharan, Gus A Baker, and Laura H Goldstein. 2008. "Psychological Treatments for Epilepsy." *Cochrane Database of Systematic Reviews*, no. 3. John Wiley & Sons, Ltd.

Raspe, Rudolf Erich. 2009. *The Surprising Adventures of Baron Munchausen.* The Floating Press.

Reips, U.D., and F. Funke. 2008. "Interval-Level Measurement with Visual Analogue Scales in Internet-Based Research: VAS Generator." Journal Article. *Behavior Research Methods* 40 (3): 699–704.

Reips, Ulf-Dietrich, and Frederik Funke. 2008. "Interval-Level Measurement with Visual Analogue Scales in Internet-Based Research: VAS Generator." *Behavior Research Methods* 40 (3). Springer: 699–704.

Ringle, C.M., and R. Schlittgen. 2007. "A Genetic Algorithm Segmentation Approach for

Uncovering and Separating Groups of Data in Pls Path Modeling." Journal Article. *PLS* 7: 75–78.

Ringle, C.M., S. Wende, and A. Will. 2010. "Finite Mixture Partial Least Squares Analysis: Methodology and Numerical Examples." Journal Article. *Handbook of Partial Least Squares*, 195–218.

Ringle, Christian M, Marko Sarstedt, and Detmar W Straub. 2012. "Editor's Comments: A Critical Look at the Use of Pls-Sem in Mis Quarterly." Journal Article. *MIS Quarterly* 36 (1): iii–xiv.

Ringle, Wende, C.M. n.d. "SmartPLS 2.0 (M3) Beta, Hamburg 2005, Http://Www.smartpls.de." Journal Article.

Roberts, Susan B, Dean M Bonnici, Andrew J Mackinnon, and Marian C Worcester. 2001. "Psychometric Evaluation of the Hospital Anxiety and Depression Scale (Hads) Among Female Cardiac Patients." *British Journal of Health Psychology* 6 (4). Wiley Online Library: 373–83.

Samuel, Mari Dominique Drouet Kotz, Dominique Drouet Mari, and Samuel Kotz. 2001. *Correlation and Dependence*. World Scientific.

Sargan, John D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica: Journal of the Econometric Society*. JSTOR, 393–415.

Schechtman, Edna, and Shlomo Yitzhaki. 1999. "On the Proper Bounds of the Gini Correlation." *Economics Letters* 63 (2). Elsevier: 133–38.

Scheines, Richard, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. 1998. "The Tetrad Project: Constraint Based Aids to Causal Model Specification." *Multivariate Behavioral Research* 33 (1). Taylor & Francis: 65–117.

Schonhoff, Thomas, and Arthur Giordano. 2006. *Detection and Estimation Theory*. Prentice Hall.

Sequeira, Henrique, Pascal Hot, Laetitia Silvert, and Sylvain Delplanque. 2009. "Electrical Autonomic Correlates of Emotion." *International Journal of Psychophysiology* 71 (1). Elsevier: 50–56.

Shevlyakov, Georgy L, and Nikita O Vilchevski. 2002. "Minimax Variance Estimation of a Correlation Coefficient for ε-Contaminated Bivariate Normal Distributions." *Statistics & Probability Letters* 57 (1). Elsevier: 91–100.

Sokal, Alan D. 1996. "Transgressing the Boundaries: Toward a Transformative Hermeneutics of Quantum Gravity." *Social Text*, no. 46/47. JSTOR: 217–52.

Sokal, Alan D, and Jean Bricmont. 1998. *Intellectual Impostures: Postmodern Philosophers' Abuse of Science*. profile books London.

Sola Pool, I. & M. Kochen de. 1979. "Contacts and Influences." *Social Networks* 1 (1). Elsevier: 5.

Sosik, John J, Surinder S Kahai, and Michael J Piovoso. 2009. "Silver Bullet or Voodoo

Statistics? A Primer for Using the Partial Least Squares Data Analytic Technique in Group and Organization Research.” *Group & Organization Management* 34 (1). SAGE Publications Sage CA: Los Angeles, CA: 5–36.

Sparks, Roland, Nick Desai, Perumal Thirumurthy, Cindy Kistenberg, and S Krishnamurthy. 2006. “Measuring E-Commerce Satisfaction: Reward Error and the Emergence of Micro-Surveys.” In *IADIS International E-Commerce Conference Proceedings*.

Sterne, Jonathan AC, and George Davey Smith. 2001. “Sifting the Evidence—what’s Wrong with Significance Tests?” *Physical Therapy* 81 (8). Oxford University Press: 1464–9.

Stigler, Stephen M. 1980. “Stigler’s Law of Eponymy.” *Transactions of the New York Academy of Sciences* 39 (1 Series II). Wiley Online Library: 147–57.

Stuart, Alan, Maurice George Kendall, and John Keith Ord. 1991. *Classical Inference and Relationship*. Oxford University Press.

Tabachnick, Barbara G, and Linda S Fidell. 2007. *Using Multivariate Statistics*. Allyn & Bacon/Pearson Education.

Tabachnick, BG, and L. Fidell. n.d. “S.(1989).” Journal Article. *Using Multivariate Statistics* 2.

Tanaka, Jeffrey S. 1987. “‘ How Big Is Big Enough?’: Sample Size and Goodness of Fit in Structural Equation Models with Latent Variables.” *Child Development*. JSTOR, 134–46.

Tao, Jianhua, and Tieniu Tan. 2005. “Affective Computing: A Review.” In *International Conference on Affective Computing and Intelligent Interaction*, 981–95. Springer.

Temme, Dirk, Henning Kreis, and Lutz Hildebrandt. 2006. “PLS Path Modeling.” Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät.

Tenenhaus, Michel, Silvano Amato, and Vincenzo Esposito Vinzi. 2004. “A Global Goodness-of-Fit Index for Pls Structural Equation Modelling.” In *Proceedings of the Xlii Sis Scientific Meeting*, 1:739–42.

Tenenhaus, Michel, Vincenzo Esposito Vinzi, Yves-Marie Chatelin, and Carlo Lauro. 2005. “PLS Path Modeling.” *Computational Statistics & Data Analysis* 48 (1). Elsevier: 159–205.

Theil, Henri. 1980. *System-Wide Explorations in International Economics, Input-Output Analysis, and Marketing Research*. Vol. 2. North-Holland.

Theil, Henri, and John CG Boot. 1962. “The Final Form of Econometric Equation Systems.” *Revue de L’Institut International de Statistique*. JSTOR, 136–52.

Thurstone, Louis Leon. 1935. “The Vectors of Mind: Multiple-Factor Analysis for the Isolation of Primary Traits.” University of Chicago Press.

Travers, Jeffrey, and Stanley Milgram. 1967. “The Small World Problem.” *Phychology Today* 1 (1). JSTOR: 61–67.

———. 1977. “An Experimental Study of the Small World Problem.” In *Social Networks*, 179–97. Elsevier.

Turner, Malcolm E, and Charles D Stevens. 1959. "The Regression Analysis of Causal Paths." *Biometrics* 15 (2). JSTOR: 236–58.

Velicer, WF, JO Prochaska, JL Fava, GJ Norman, and CA Redding. 1998. "Detailed Overview of the Transtheoretical Model." Journal Article. *Homeostasis* 38: 216–33.

Vinzi, Vincenzo Esposito, Carlo N Lauro, and Silvano Amato. 2005. "PLS Typological Regression: Algorithmic, Classification and Validation Issues." In *New Developments in Classification and Data Analysis*, 133–40. Springer.

Watts, Duncan J, and Steven H Strogatz. 1998. "Collective Dynamics of 'Small-World'networks." *Nature* 393 (6684). Nature Publishing Group: 440.

Wedderburn, Robert WM. 1974. "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss—Newton Method." *Biometrika* 61 (3). Oxford University Press: 439–47.

Wehrens, Ron, and B-H Mevik. 2007. "The Pls Package: Principal Component and Partial Least Squares Regression in R."

Werts, Charles E, Karl G Jöreskog, and Robert L Linn. 1972. "A Multitrait-Multimethod Model for Studying Growth." *Educational and Psychological Measurement* 32 (3). Sage Publications Sage CA: Thousand Oaks, CA: 655–78.

Werts, Charles E, Robert L Linn, and Karl G Jöreskog. 1974. "Intraclass Reliability Estimates: Testing Structural Assumptions." *Educational and Psychological Measurement* 34 (1). Sage Publications Sage CA: Thousand Oaks, CA: 25–33.

Westland, J Christopher. 2010. "Lower Bounds on Sample Size in Structural Equation Modeling." *Electronic Commerce Research and Applications* 9 (6). Elsevier: 476–87.

———. 2011a. "Affective Data Acquisition Technologies in Survey Research." *Information Technology and Management* 12 (4). Springer: 387–408.

———. 2011b. "Electrodermal Response in Gaming." *Journal of Computer Networks and Communications* 2011. Hindawi.

Westland, J.C. 2010. "Lower Bounds on Sample Size in Structural Equation Modeling." Journal Article. *Electronic Commerce Research and Applications* 9 (6): 476–87.

Wildt, Albert R, and Michael B Mazis. 1978. "Determinants of Scale Response: Label Versus Position." *Journal of Marketing Research.* JSTOR, 261–67.

Wittfogel, Karl A. 1957. "Chinese Society: An Historical Survey." *The Journal of Asian Studies* 16 (3). Cambridge University Press: 343–64.

Wold, H. 1966. "Estimation of Principal Components and Related Models by Iterative Least Squares." Journal Article. *Multivariate Analysis* 1: 391–420.

———. 1974. "Causal Flows with Latent Variables: Partings of the Ways in the Light of

Nipals Modelling." Journal Article. *European Economic Review* 5 (1): 67–86.

Wolfram, Stephen. 2002a. *A New Kind of Science*. Book. Vol. 5. Wolfram media Champaign.

———. 2002b. *A New Kind of Science*. Vol. 5. Wolfram media Champaign, IL.

Worcester, Robert M, and Timothy R Burns. 1975. "Statistical Examination of Relative Precision of Verbal Scales." *Journal of the Market Research Society* 17 (3). Market Research Society: 181–97.

Wright, S. 1921. "Correlation and Causation." Journal Article. *Journal of Agricultural Research* 20 (7): 557–85.

———. 1934. "The Method of Path Coefficients." Journal Article. *The Annals of Mathematical Statistics* 5 (3): 161–215.

Wright, Sewall. 1960. "Path Coefficients and Path Regressions: Alternative or Complementary Concepts?" *Biometrics* 16 (2). JSTOR: 189–202.

Xu, Weichao, YS Hung, Mahesan Niranjan, and Minfen Shen. 2010. "Asymptotic Mean and Variance of Gini Correlation for Bivariate Normal Samples." *IEEE Transactions on Signal Processing* 58 (2). IEEE: 522–34.

Ye, H Ge, X, and WL Shiau. 2010. "Website Quality and Consumer's Purchase Intention? Product Information, Navigation, and Visual Aesthetics." working paper.

Zellner, A. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." Journal Article. *Journal of the American Statistical Association*, 348–68.

Zellner, A., and H. Theil. 1962. "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations." Journal Article. *Econometrica: Journal of the Econometric Society*, 54–78.

Zellner, Arnold, and Henri Theil. 1962. "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations." *Econometrica: Journal of the Econometric Society*. JSTOR, 54–78.