Clemens Pirker

Statistical Noise or Valuable Information

The Role of Extreme Cases in Marketing Research

GABLER

RESEARCH

Clemens Pirker

Statistical Noise or Valuable Information

GABLER RESEARCH

Clemens Pirker

Statistical Noise or Valuable Information

The Role of Extreme Cases in Marketing Research

With a foreword by Prof. Dr. Hans Mühlbacher

GABLER

RESEARCH

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Dissertation University of Innsbruck, 2008

1st Edition 2009

Editorial Office: Claudia Jeske | Nicole Schweitzer

Gabler is part of the specialist publishing group Springer Science+Business Media.
www.gabler.de

Cover design: KünkelLopka Medienentwicklung, Heidelberg
Printed on acid-free paper
Printed in Germany

ISBN 978-3-8349-1762-1

Foreword

Contemporary marketing research is dominated by empirical studies and quantitative methods for data analysis. In quantitative studies, researchers frequently face the problem of a relatively small number of subjects deviating substantially from the remaining observations. The question arises of how to proceed: Are these deviating observations errors, which need to be corrected, or phenomena worthy of closer investigation?

Clemens Pirker's dissertation extensively deals with this question. He documents how published marketing research in leading journals approaches the problem in very pragmatic ways, often simply eliminating substantially deviating cases from the data set or correcting the data by statistical means. Such approaches increase the statistical fit of models and help verify hypotheses. In view of widespread verificationist research, this "easing" approach is not surprising. But, potentially interesting information contained in deviating data are neglected. An opportunity for scientific progress may be lost. The author presents alternative ways of dealing with outliers, based on a sound discussion from a philosophy of science point of view. He concludes his work with an exhaustive empirical study showing how deviating data may help gaining new insights, not only relevant for scientific progress but also relevant for management decisions.

This dissertation offers an interesting repertory of suggestions for all readers interested in scientific progress through empirical studies who are - at the same time - open for sound methodological discussions based on philosophy of science. This work should be obligatory reading in the training of starting academics.

Univ. Prof. Dr. Hans Mühlbacher

Preface

"When nature presents us with a group of objects of every kind, it is using rather a bold metaphor to speak in this case also of a law or error." John Venn (1834 – 1923)

This book deals with extreme and outlying cases – a topic which looks back on several centuries of scientific research. Thanks to Malcom Gladwell's "Outliers" and Nassim Nicholas Taleb's "The Black Swan", it recently received unknown popularity and made it to the best – seller lists. The book at hand roots in more common circumstances: The discovery of outlying cases while analysing an extensive dataset marked the starting point. In the same way as the above mentioned books, it does not aim for being exhaustive on all possible fields where these cases are relevant. It aims to inspire and to propose an alternative view on the phenomenon in the field of Marketing Research without ignoring the bigger picture and importance in other fields. The book at hand represents a slightly modified version of my dissertation I completed in early 2008.

A significant number of people have contributed to making this dissertation become real: I am grateful to my supervisor Prof. Dr. Hans Mühlbacher, head of the Institute of Strategic Management, Marketing and Tourism at the University of Innsbruck for his support and sponsoring of my doctorate journey. I am particularly indebted to him for showing me a different perspective on Marketing, countless inspiring and challenging thoughts and for giving me an academic home at the department

In the same way, I am grateful to my co-supervisor Prof. Arch G. Woodside PhD who supported me with invaluable advice throughout the whole process and provided data for the empirical study. His invitations to research visits at Boston College and the University of South Carolina broadened my horizon substantially.

I owe much gratitude to the Marketing team and the doctoral students at the Department of Strategic Management, Marketing and Tourism. I am particularly grateful to Prof. Dr. Andrea Hemetsberger for supporting me in writing my first academic papers, to Dr. Rudi Massimo for mentoring me in my first steps at the department and to Pauline Fusseis for supporting me in uncountable ways.

A special thanks goes to Prof. Dr. Betty Thorne and Dr. Reinhard Prügl for their invaluable comments and suggestions for the preparation of the final version of this dissertation as well as to Prof. Dr. Mike Etzel and Prof. Dr. Stefan Lang for their comments during the project.

Last but first, my deepest gratitude goes to my parents Ingeborg and Harald, my sister Marie-Theres, my brothers Dominik and Benedikt and my girlfriend Edith, who all encouraged me in the process and offered me much needed, unlimited support during my doctorate. My gratitude to them is boundless. I would like to dedicate this book to them with love.

Clemens Pirker

Table of Contents

Index of Tables

Index of Figures

Index of Exhibits

A. Introduction

Frequently, marketing academics and market researchers encounter extreme and deviant cases in their studies. These cases highly influence statistical analyses while researchers try to maximize representativity and external validity. Thus, many opt to exclude, to average out or to change the value of these cases in order not to disturb their results. This phenomenon is not limited to marketing data as the following example shows.

Chlorofluorocarbons (CFC) – which were intensively used in aerosols, refrigerators and air conditioning appliances – are today claimed to substantially increase the hole in the ozone layer of the Earth. In the early 1970s, their usage flourished while researchers became anxious as laboratory experiments unveiled first evidence of the destructive effect of CFCs on the ozone layer. This made many countries such as France, Great Britain, Japan, the Soviet Union and the United States launch scientific projects to record and analyse ozone data in Antarctica. Besides their ground based stations, NASA launched the Nimbus 7 satellite programme in 1978 to measure the ozone layer. But neither the ground based instruments nor the satellite measurements indicated any evidence for a depletion of the ozone layer.

Surprisingly, a conference paper by the Japanese Antarctic Research team in 1984 reported very low ozone values from the previous year (Chubachi 1984). Even more important, a "Nature" article in 1985 by the British team revealed a 40% depletion of the Ozone layer protecting the Earth (see Farman, Gardiner and Shanklin 1985).

The scientific community was stunned and made other teams concerned about their data recording. Other projects were investigated for their scientific practices resulting in considerable insights: When the Nimbus 7 satellite team "reworked" their data, they discovered similarly low values. Unfortunately, the software had been programmed to "flag" and set aside values that deviated greatly from the expected measurements and so the initial measurements that should have set off the alarms were overlooked. Thus, scientists failed to detect Ozone depletion years earlier because it was much more severe than expected. The "reworked" findings were published in the 1986 "Nature" article of the team (Stolarski, Krueger, Schoeberl, McPeters, Newman and Alpert 1986). British and American teams had to admit that a lag in time of more than five years had occurred due to uncritical data analysis and particularly outlier exclusion combined with a too narrow view on the phenomenon. Finally, the British and the NASA publications on ozone depletion made 43 nations sign the Montreal protocol for the reduction of CFCs in 1987 (see Maureen 2001 for a more detailed account).

Outliers are not only relevant to natural sciences. Extreme and deviant cases have contributed in various other domains to help humans to gain knowledge. Going back in the history of science, the study of extremes is characterized by a long academic tradition. For instance, Greek philosophers like Anaximander (610 – 545 B.C.) built their theories about the world on extremes like earth and sky and their interconnection. Hegel's (1770 - 1831) dialectic logic can also be seen from this perspective. He takes an idea (thesis), entertains its opposite (antithesis) and attempts to fuse these into a third idea (synthesis) (Barett 1997). Many times, extreme and deviant cases are considered abnormal. Scientific theories such as Copernican astronomy or continental drift have advanced in spite of apparent refutations by empirical data. Or, as Lakatos (1974, p135) puts it: "at one time or another, [major scientific discoveries] were all in danger of drowning in an 'ocean of anomalies' ".

After concentrating on general decision modelling and major trends, marketing has also discovered extreme and "micro" phenomena (Hunt 1994) as a relevant area for research. Many of those extreme phenomena have been recognized by leading academics and also partly found their way in to business applications. People exhibiting extreme and deviant behaviour compared to the main stream often create trends that are adopted by other consumers (such as lead users in Von Hippel 1988), they sometimes develop products themselves which are later produced by companies (e.g. Kotler 1986). Others show extreme forms of relationships, such as attachment (e.g. Wallendorf and Arnould 1988), or devotion to products or brands (Pimentel and Reynolds 2004). To make a long story short, extreme and deviating cases are frequently the focus of scientific investigations.

This dissertation aims to shed light on the role of extreme cases in recent marketing literature and the practices related to their investigation and treatment. It investigates the role of the philosophical perspective as well as their potential contribution to theory development.

1. Problem statement and research questions

Marketing and other social science researchers frequently encounter extreme and outlying cases when they proceed from data collection to analysis. Oftentimes, they cannot clarify whether these unusual values should be considered as error-laden observations or valuable information.

(1) “Do we lose important information if we continue skipping them?” and, if yes “What should marketing research do about it?”

…may be the most striking questions. The subsequent situation can be described as somewhat of an unresolved dilemma. On the one hand, these cases may disturb analysis procedures and consequently impact results such as fit indices, coefficients, or simple means and standard deviations. Every researcher in his role as an author is aware that such distortion may highly impact the probability of the results being accepted for publication. On the other hand, research practices suggested by – for example – the publication manual of the American Psychological Association (APA) or publications by senior researchers (e.g. Finney 2006) warn that “naïve rules on rejecting observations may have ethical implications”. This is also related to additional information covered in theses cases, which may be lost if a too narrow focus exists on a phenomenon. The recommendations of some statistics instructors to proceed with deleting outlying observations (e.g. Glaister and Glaister 2004) only worsen the dilemma. This dissertation therefore addresses the following sub-questions:

(2) How are extreme and outlying cases handled in the marketing science literature? What are the typical approaches? What problems arise with these approaches?

These questions are highly interrelated with another aspect of research: the methodological foundations. Similar to many industry practices, marketing research is also characterized by “schools of thought” emerging from different research paradigms with different assumptions. Even though Hunt (1994) declared Critical Realism as the “winner” of the philosophical debate, these paradigms are inherent in every study and influence which methods are used and their way of implementation. Therefore…

(3) How do different philosophical schools view outlying and deviant cases occurring in statistical analysis? What do they recommend subsequently?

The effect of dealing with extreme and outlying cases in different ways has rarely been studied in marketing literature. Only Mullen, Milne and Doney (1995) investigated the effect of excluding and retaining outliers within a structural equation model. The study of this dissertation aims to investigate this question along the methodology of a published study.

(4) What is the impact of different approaches to outlier handling on the results of a scientific study?

Finally, the in-depth investigation of cases which deviate from the main trend have been advocated to increase precision in social science (Chase 2001, Julander and Söderlund 2002) by e.g. analysing extreme groups separately (Mahajan, Sharma and Wind 1984). It is well known that an analysis on group level often reveals different results than a study of the whole population (Bass, Tigert and Lonsdale 1968; Sharma, Durand and Gur-Arie 1981). In many marketing related domains, extreme and deviant cases have stimulated research that yielded fruitful theoretical insights: Such research includes lead users (Von Hippel 1988), consumer co-creation (e.g. Kotler 1986) or criminal incidents as in Lin and Brown (2006), though not always labelled as “outliers”. While Julander and Söderlund (2002) have initially probed the strategy of identifying additional theoretical insights through outliers on a customer satisfactions experiment, this dissertation intends to extended on their work.

(5) How can extreme and outlying cases contribute to the theoretical knowledge of a scientific field?

This dissertation’s ultimate goal is to identify an epistemologically sound way of using outlying and deviant cases instead of excluding them in a blindfolded way.

2. Research approach

The research approach of this dissertation follows Phillips and Pugh's (1994) three characteristics which contribute to good research and distinguish a dissertation project from activities such as decision making or consultancy. These include:

- An "open system of thought" which requires "continual testing, review and criticism of others' ideas and a willingness to hazard new ideas, even if one can't find half a dozen references to support one's view".
- Critical data examination, as well as requesting evidence drawn from the conclusions.
- The attempt to always generalize the research, but within stated limits: to extract understanding from one situation and to apply it to as many other situations as possible (Easterby-Smith and Thorpe 2002, p19).

Researchers are influenced by the paradigms they represent i.e. their sets of beliefs combined with methods. At the base of each scientific study lie the scientific assumptions by the researcher. It is widely recommended that researchers declare these underlying assumptions. (Miles and Huberman 1994). Nevertheless, many publications do not discuss their assumptions in an explicit way. As conceptualizations are representations of the real world, the researcher's view of the world should guide the way the study is carried out. If not stated explicitly in a particular section, this dissertation takes a critical realist approach.

3. Dissertation structure

In section A, this dissertation introduces the research questions and the scientific approach as well as the underlying paradigm.

Section B first studies extremes and outliers from the perspective of existing phenomena in different scientific disciplines, which in the case of data collection, manifest as extreme cases. This part also shows the relevance of this research topic to different domains. Part two of section B summarizes the literature in different domains, and treats this question on a per case basis. Statistics and datamining literature quite naturally represent a substantial part of this section. This section answers research question (RQ) 1 from a theoretical point of view.

Section C studies the practices in marketing research related to extreme cases. As practices are particularly influenced by the researcher's view of the world, research paradigms are presented first (RQ 3), while the respective handling of the cases follows (RQ 2) which is also related to the creation of new knowledge (RQ 5). Then, the handling of outliers is investigated in a publication audit and quantified (RQ 2). Part three places outlier-handling practices in the larger context of current criticism of research practices.

Section D studies the consequences of the most "popular" handling strategy by replicating a published study that has applied this strategy. Thereby, the consequences are shown (RQ 1) and an alternative approach is probed with the goal of improving current practices (RQs 4 and 5).

Section E summarizes the results, highlights the contribution to the literature, outlines the limitations of the work and draws implications for research and practice. Finally, further fields of study are discussed.

B. Literature Review

This chapter approaches scientific literature from two different ways: extreme and exceptional phenomena (1) as well as extreme and exceptional cases (2).

Section one of this chapter examines literature starting from a scientific argument or an established theory before approaching data. This examination includes a wide variety of disciplines where empirical data are used. For their empirical investigations, such publications purposefully select a small group of people or small amount of observations that behave in an extreme or exceptional manner in comparison to a larger group of people or observations. Section two depicts literature that departs from given data and investigates how to deal with these exceptional cases and their sources. Particularly statistics and data mining research have shown interest in this area, but also social sciences to a lesser extent. This approach can also be called a “data driven perspective”.

1. Extreme and exceptional phenomena

While many fields of science have studied averages and main streams such as median income, average rainfall or average risk of infections of diseases, a significant number of studies also deal with extremes and exceptionality. This research includes estimating maxima or minima on some kind of scale, the investigation of cases exhibiting extreme values or the detailed analysis of groups showing exceptional behaviour.

1.1. Engineering, natural science and medicine

In medicine, extremes play an important role: Watson and Crick’s discovery that the DNA consists of two identical strands of molecules connected together in opposing spirals is claimed to be the result of contemplating two or more opposites at the same time (Rothenberg 1979). Another example is the investigation of rarely occurring dysfunctions, such as “Down Syndrome”, which would never have reached enough attention by applying widely used significance criteria in social science.

Design values of engineering works such as dams, buildings or bridges are obtained on a compromise between safety and cost, between guaranteeing that they survive when subject to extreme operating conditions and budgetary constraints. Estimating extreme capacities or operating conditions is very difficult because of the lack of available data. The use of safety factors has been a classical solution to the problem. Recent research shows that this approach is not completely satisfactory in terms of safety and cost, because high probabilities of failure can be obtained on one hand and large and unnecessary waste of money on the other. Knowledge of the distributions of maxima and minima of the relevant phenomena is important in obtaining good solutions to engineering design problems (Castillo, Hadi, Balakrishnan and Srabia 2005).

In the field of ocean engineering including e.g. harbour constructions or offshore platforms, engineers rely on knowledge about the distributions of wave heights and periods. While the average waves are interesting, the engineer focuses on the periods associated with the largest waves. Some recent publications in this field include Sjö (2001) and Onorato, Osborne and Serio (2002).

Structural engineering has to deal with modern building codes. Standards provide information on (a) extreme winds in the form of wind speeds corresponding to various specified mean recurrence intervals, (b) design loads, and (c) seismic incidence in the form of areas of equal risk. Wind speeds are estimates of the maximum strength that can occur where the building or engineering work is located. These estimates impact design and costs significantly. Design loads relate to the largest loads acting on the structure during its lifetime and can influence whether the structure collapses or money is wasted if overestimated. Furthermore, the building will only survive if it resists the most severe earthquake occurring during its design period. Precise estimations of the probabilities of occurrence of extreme winds, loads and earthquakes are required in order to allow for realistic safety margins in structural design on one hand, and for economic solutions on the other (see Castillo, Hadi et al. 2005 p5 for a literature overview).

The reoccurrence of extreme hydrological events such as floods and droughts is important in reservoir storage and yield planning, drought studies and reservoir operation. The storage capacity for a headwater – for instance – is frequently based on a critical historical drought sequence. To build dams and canals, stream discharge and flood flow have long been measured and used by engineers in the design process. Especially since floods have caused high losses in human life and property, the need for more precise estimates of these extremes has been promoted. The design of these hydraulic structures is not only affected by the intensity of floods, but also by their frequency of occurrence. A more accurate estimation of both data can therefore improve floodplain management and risk assessment (e.g. Morrison and Smith 2001, Karr 1976).

Meteorological conditions can significantly influence aspects of human life and nature. Oftentimes, scientists respect mean values, but focus their interest on the occurrence of extreme events, such as very low temperatures or high levels of rainfalls. The accurate prediction of these rare events is therefore crucial (see discussion by e.g. Ferro and Segers 2003, Galambos and Macri 2002, Leadbetter, Lindgren and Rootzén 1983 and Sneyers 1984).

Extreme value theory is applied to material strength for the analysis of size effects. Many times, the strength of actual structures is estimated from the strength of small elements of reduced size, samples, prototypes or models that are tested under laboratory conditions. Extrapolation from these small sizes to the size of the final application is needed. Extreme value theory helps to analyse the size effect and make these extrapolations reliable. The weakest element determines the strength of the entire structure. Thus, large pieces are statistically weaker than small pieces (see e.g. Harter 1977, Harter 1978 for a literature overview). Similarly, corrosion failure is caused by a chemical process originated from the progressive size increase and penetration of initially small pits through the element. The corrosion resistance of an element is determined by the largest pits and largest concentrations of chemical agents and that small and intermediate pits and concentrations do not have any effect on the corrosion strength of the element (see e.g. Reiss and Thomas 2001, Thiruvengadam 1972).

Due to economic considerations, many highways are designed in a way that traffic collapses only a limited number of times. As a point of departure for the calculation, the n^{th} largest traffic intensity during the particular period is taken. If reliable probability distributions are available, the theory of extreme order statistics allows a reliable design to be produced (e.g. Glynn and Whitt 1995 or Gómez-Corral 2001)

Large concentrations of people – or the appearance of new industrial settlements – are often reflected in the pollution of air, rivers, coasts and natural environment, in general. Government regulations attempt to maintain pollution levels below a certain critical level. These regulations are only met if the largest pollution concentration remained below this critical level. The largest value therefore plays a critical role in the design.

Error free measurement, analysis and prediction of extreme values are central to many applications in engineering, natural science and medicine. A detailed investigation of extreme phenomena and their generating mechanisms allows a better understanding and supports the development of measures helping to prevent floods or construction errors, for instance.

1.2. Sociology and social psychology

In psychology and sociology, many extreme and exceptional phenomena can be found in the area of deviant behaviour. One of the most recent overviews by Thio (2006) covers topics such as physical violence, rape and child molesting, family violence, suicide, mental disorders, heterosexual deviance, homosexuality and other victims of stigma, legal and illegal drug use, drinking and alcoholism and privileged vs. underprivileged deviance.

Deviance can be positive and negative. Committing a deviant act does not make a person a deviant, especially when the act is kept secret and unlabeled by others as deviant (Thio 2006). Some even go beyond this notion. They argue that people who hold the power to label, either by themselves, by influencing public opinion or both, should be at the centre of interest, irrespective of whether they use this power or not (e.g. Simon 2002, Ermann and Lundman 2002).

In terms of defining deviance, a philosophical split between positivist and constructionist research is frequently outlined: While the positivist defines deviance as observable, determined and real behaviour, a constructionist would consider it more often as a social construction – an idea imputed by society to some behaviour. Consequently, the objects of study, methods and theories vary. The focus of positivists is on deviant behaviour and deviant persons rather than non-deviants who label others deviants. This latter category includes lawmakers and law-enforcers, which would be more likely studied by constructionists.

Positivists also argue that deviance is determined by forces beyond the individual's control; hence, causes are important to study. Constructionists focus more on the understanding of deviants. They advocate a subjective approach and frequently present deviants as conventional people. Deviants are not presented as if they were robots, passively and senselessly developing a poor self-image as conventional society expects of them. Rather, they are described as actively seeking positive meanings in their deviant activities. In Jack Katz's (1988) analysis, for example, murderers see themselves as morally superior to their victims, which gives them dignity and respectability because their victims have unjustly humiliated them by taunting or insulting them. The research focus of constructivist researchers lies more on how social agencies define certain people as deviant and carry out sanctions against them. Recapitulating, the constructionist perspective sees deviant behaviour not as real in and of itself, but as a label. Because of its subjective nature, it needs to be studied with methods allowing subjectivity and empathy. Deviant behaviour is voluntary and a self-willed act rather than one caused by forces in the internal and external environment (Thio 2006).

Table 1 and Table 2 aim to summarize the major theories from positivist and constructivist perspective that are used to explain deviance.

Positivist Theories of Deviance
Anomie – Strain Theory: Social strain causes deviance. Anomie is conceived as the breakdown of social norms that results from the society's urging of people to be ambitious while failing to provide them with the opportunities to succeed. ▪ Merton's goal-means gap: Deviance is prevalent in society because the society encourages people to achieve success without providing equal opportunity for achieving it (Merton 1938). ▪ Cohen's status frustration: Deviance is prevalent among lower-class youths because they fail to achieve status in a middle-class school environment (Cohen 1955). ▪ Cloward and Ohlin's differential illegitimate opportunity: Lower-class youths are likely to engage in delinquent activities if they have access to illegitimate opportunity (Cloward and Ohlin 1960). ▪ Latest version of the theory: The American Dream contributes to deviance by directly encouraging the use of illegal means to achieve success, while various social strains cause deviance such as the emotions frustration and anger (Agnew 1992, Messner and Rosenfeld 2001). These theories have been criticized for not having reliable evidence and using official statistics, thus failing to capture the whole deviance picture as higher class people are much less likely to be caught by law enforcers for example. Secondly, empirical studies have shown that lower class people hold lower levels of success aspirations in contrary to the equal level in all classes as the theory would assume.
Social Learning Theory: Deviant behaviour is learned through social interaction ▪ Sutherland's differential association: People are likely to become deviant if they associate with people holding deviant ideas rather than with people holding anti deviant ideas (Sutherland 1939). ▪ Glaser's differential identification: People are likely to become deviant if they identify themselves more with deviants than with non-deviants (Glaser 1956). ▪ Burgess and Aker's differential reinforcement: Deviants are likely to continue engaging in deviant activities if they have been rewarded rather than punished for their past deviance (Burgess and Akers 1966). Questions about the empirical interpretation of Sutherland's differential association have been raised, even though many researchers support the idea through their own data interpretation. No conclusive evidence has been found for Glaser's theory that identification with delinquent friends is the cause or antecedent of delinquency. Burgess and Aker's theory suffers from lacking to explain the initial deviance act, while the subsequent ones are covered.
Control Theory: A large body of literature has accounted for the absence of social control as determinant for deviant behaviour. Some prominent theories are as follows: ▪ Hirschi's social bond and self-control: People are likely to become deviant if their bond to society and their self-control are weak (Hirschi and Gottfredson 1994; Hirschi and Gottfredson 2000) ▪ Braithwaite's re-integrative shaming: People are likely to become deviant if they are not made to feel ashamed for their wrongdoing or to feel they are an integral part of society (Braithwaite 1989, 2000). ▪ The deterrence doctrine: People are likely to become deviant if they know their deviant acts are not punished with severity, certainty, or swiftness (see. e.g. Thio 2006). Great support has been raised for Hirschi's social bond and self control theory. Nevertheless some of the studies use cross - sectional data which is not ideal to evaluate longitudinal bonds. Others use deviant acts as indicators of weak self control, i.e. they suggest that deviance causes deviance. Braithwaite's theory of shaming is claimed to work only for those people who still retain a sense of shame for their crimes. Empirical studies nevertheless question the deterrence doctrine as more severe punishment proved the likeliness of committing crimes again (e.g. Wright, Caspi, Moffitt and Paternoster 2004). Other criticism includes that social control should also be seen as a cause of deviance, not only a preventer. Source: Table compiled and adapted from Thio (2009)

Table 1: Positivist theories of deviant behaviour

Besides positivist theories, constructionist approaches have gained high acceptance in the field of deviant behaviour. The most prominent is still labelling theory, which is actually a version of symbolic interactionism, a well-accepted sociological paradigm dealing with social behaviour in general.

Constructionist Theories of Deviance
Labelling Theory: Relatively powerful people are more likely to label the less powerful as deviant than vice versa, and being labelled deviant by society leads people to see themselves as deviant and live up to this self-image by engaging in more deviancy (e.g. Erikson 1962). Frequent criticism includes that this theory does not provide a cause for deviance, even though it does not set out for this aim. Some studies succeeded in showing that labelling encouraged further deviance, but it often simply did not. At last, the theory cannot deal with different degrees of visibility of deviance (i.e. hidden deviance would not be considered as deviance).
Phenomenological Theories: Looking into people's subjective interpretation of their own experiences is key to understanding their deviant behaviour (e.g. Morris 1977; Handel 1982; Roubach 2004). It is widely accepted that the subjective approach through e.g. ethnographic studies has brought new understanding for deviant behaviour. However, this approach is in conflict with the positivist emphasis on the objective side of deviant reality.
Conflict theory: ▪ Legal Reality: Law enforcement officials favour the rich and powerful over the poor and weak (Chambliss 1969; Chambliss and Seidman 1971). ▪ Social Reality: The dominant class produces crime by making laws, enforcing laws, oppressing subordinate classes, and spreading crime ideology (Quinney 1974, Quinney 1975). ▪ Feminist: Conventional theories of deviance are largely inapplicable to women, and the status of women as victims and offenders reflects the continuing subordination of women in a patriarchal society. ▪ Postmodernist: Privileged language of the powerful dominates the marginalized language, and thus the lives, of the weak as deviants (e.g. Arrigo and Bernard 1997). Conflict theory contributes to the understanding of how social inequality, such as in the form of capitalism and patriarchy, influences the making and enforcing of norms, rules or laws or the definition, production and treatment of deviance in society, but also the motivation behind lawmaking. Source: Table adapted and compiled from Thio (2009)

Table 2: Constructionist theories of deviant behaviour

Today's research on deviant behaviour is much more interwoven and makes a clear split difficult. A more recent definition of deviant behaviour is "any behaviour considered deviant by public consensus, which may range from the maximum to the minimum" (Thio 2006, p23).

Besides many mechanisms generating extreme and deviant observations, sociology and social psychology research asks for the influence of the scientific perspective on the treatment of outliers. Furthermore, it probes whether they are objective or subjective (i.e. researcher determined) phenomena. In the second case, labelling theory may help to explain treatment practices.

1.3. Marketing research

Marketing and related disciplines have integrated extreme and exceptional phenomena in the core of their knowledge to a lesser extent than many other scientific disciplines. Nevertheless, the following examples illustrate that many areas of marketing science have also researched extreme and exceptional phenomena.

In the area of new product development, Von Hippel (1988) shows that lead users can be a valuable source of new ideas in product development. They "…face needs that will be general in a marketplace, but they face them months or years before the bulk of that marketplace encounters them" and "… are positioned to benefit significantly by obtaining a solution to those needs"(p107). They can serve as a need forecasting laboratory in market research and help to develop products more efficiently. Nevertheless, they often only represent a small part of the total users (e.g. 18% of the users of OPAC Software in Morrison, Roberts and Von Hippel 1999) and experience new or different needs than other users, which are often unfulfilled (Urban and Von Hippel 1988).

In advertising, the "tipping point" by Malcolm Gladwell (2000) explains "social epidemics" which start with very few people and may spread at an epidemic speed. From various examples, he derives three types of people who promote social epidemics but are rare and exceptional in society. Connectors have a broad social circle, often across social classes and link those social circles. Mavens represent knowledgeable people who discover e.g. market asymmetries through their depth of knowledge. Salesmen are people with powerful negotiation skills who influence other people - often in a soft and subconscious way. While the concept is originally rooted in sociology, Gladwell applies it to such different fields as teenage smoking, criminality (where single individuals can start a major wave) and advertising, where the ability to generate such epidemics is highly sought after. In the same way, he explains that for companies such as Airwalk, choosing the "right" advertising agency can also start a sales growth with epidemic dimensions (p200).

The concept of microtrends, promoted by Mark Penn (2007), is based on the idea that the most powerful forces in the society are emerging, counterintuitive trends that shape tomorrow. Microtargeting, i.e. identifying small intense subgroups and communicating with them about individual needs and wants is to become critical. Sports, for instance, may change from communal to individual rites. America's fastest growing sports are archery, skateboarding, kayaking, snowboarding, backpacking, mountain biking and bow-and-arrow hunting. Another example lies in the US workforce, where 5 million people aged 65 or older work. This number is expected to increase as the baby boomers enter that stage in life. Possible consequences include "squeeze on younger employees, who have been waiting their turn to take the reins" (p31) or the avoidance of the predicted crisis in social security systems.

In consumer behaviour, people exhibiting strong relationships to objects (e.g. Wallendorf and Arnould 1988) have been studied. Certain objects can help people to express themselves and serve as symbols for strong interpersonal ties. In their study on product attachment and possessiveness, Wallendorf and Arnould (1988b) measured the liking shown by some users. They unveiled some extremes in the US-American participants. While the liking of objects (most of them chose one linked to personal history) gradually increased until the age group of 55 – 64, the score afterwards dropped heavily and, therefore, deviated strongly from the trend. This suggests that self selected objects lose importance when people reach a higher age – an effect which has been investigated by Belk, who also suggests that material objects have highest importance in mid-life (Belk 1986). Observations of extreme product relationships are consistent with data from Sherman (1977). Pimentel and Reynolds (2004, p1) observed extremely intense loyalty to brands which they labelled consumer devotion and which "survives poor product performance, scandal, bad publicity, high prices and absence of promotional efforts".

In the area of sales management, "Heavy half" research has a long tradition. Twedt (1964) investigated different purchasing categories of the Chicago Tribune by cutting them into a light (purchasing) half and heavy (purchasing) half. He found that one heavy half household equals nine light half households, on average, and that “the heavy using household buys more, more often and more different brands”. But he notes that these households are “not identified in terms of other characteristics” (p71). Twenty years later, Cook and Mindak (1984) remarked on “a striking degree of stability, even for the national sample”. Nevertheless, demographics poorly describe reasons for this disproportionate product use, as noted by Clancy and Shulman (1994). Even if a researcher succeeds in forming socio-demographic profiles, these profiles could be purely incidental, just descriptive but not explanatory. Today, sales and retailing literature still recommends that the best way to increase sales of a product is to sell more to existing customers (Finkleman 1974; Underhill 1999 in Goldsmith 2000), so heavy users are the most likely prospects for a sales pitch (Thomas and Kirchner 1991). Further ideas in this direction suggest that marketers need to get close to their best customers, the ones most loyal to their brand.

Extreme and exceptional phenomena can also be found in the bordering disciplines of organizational science and strategy research. Weak signals, for instance, have been successfully used in both disciplines to indicate change. “High Reliability Organizations” such as aircraft carriers have established processes so that the organization is receptive and can learn from even the smallest of errors. In a particular case, a screw lost by a mechanic on the flight deck would involve the stopping of all take-offs and landings of combat planes as one of the jet engines could absorb the screw and consequently explode, something which had happened in the past (Weick and Sutcliffe 2001). They are aware that “small moments of inattention and misperception can escalate into serious adverse events” (p49). Therefore, they take into account false alarms for this high level of receptiveness in order to also address the highest possible number of problematic situations. Weick and Sutcliffe (2001) argue that the conscious search for such weak signals (in this case counter evidence) antagonizes the human tendency to declare something abnormal as normal in order to confirm ones proper expectations. They explain that often large amounts of data are overlooked which would indicate the emergence of problems. Weick and Sutcliffe (2001) summarize that many organizations could benefit from studying errors and cases which do not confirm the mainstream thinking in the organization.

Overall, marketing research and neighbouring disciplines have also paid attention to extreme and deviant phenomena using a narrow focus.

1.4. Communalities

Researchers in various disciplines have investigated people, objects and events exhibiting extreme and deviant behaviours or characteristics. In engineering, they play an important role in many applications indispensable in today's life. These cases are likely to contain important information that can provide additional insights and increase the precision of the analysis and subsequent decisions. The knowledge of whether a deviant or extreme case arises from error or not is highly relevant for many of the applications discussed. In some cases, researchers, engineers or decision makers pay a high premium – such as false alarms on an aircraft carrier – to detect the rare events which are really dangerous, in order not to miss a single one. Section 0 addresses the discussion on the sources of extremeness, including errors, in more detail.

This section has also shed light on the attitude towards such a case: More precisely, whether one considers extremeness or deviance an absolute and thus an objective or a subjective phenomenon. From a subjective perspective, labelling (i.e. the power of determining what is deviant or extreme) is worthwhile remembering. Labelling nevertheless does not only apply to deviant phenomena: Those who analyse data in research departments and decide how to present it also have the power to label cases as to whether they are e.g. extreme, outlying, and deviant or not. After an experiment on the subjective nature of outlier "labelling", Collet and Lewis (1976) conclude that "…an individual's willingness to perceive an outlier depend on the method of presentation (random, ordered, or graphical), on experience, and on the scale of the data; extreme observations tend to appear more discrepant as the scale is increased" (in: Beckman and Cook 1983, p121). Labelling is particularly relevant when such cases arise unintentionally through data collection. Section 0 investigates the way statistical literature addresses the question of which cases are to be labelled as extreme and outlying. Taking an objective stance, clear criteria for the determination of such observations are important. Section 0 also aims to study whether such criteria exist. Additionally, the question of objectivity and subjectivity in research also involves a discussion of the epistemological foundations. Researchers are influenced by those paradigms when conducting their work and making decision. Thus, they are addressed in C.1.

One can conclude that extreme and exceptional phenomena have contributed substantially to many scientific disciplines, including marketing. Nevertheless, respected researchers also admit that whether a case deviates from the main-stream lies more often with the researcher than in the field data (Shugan 2006).

2. Exceptional cases – data driven perspective

Extremeness and deviance can also be addressed from a data driven perspective. While statistics literature has shaped the foundations and developed extensive knowledge, it will be taken as a reference for the general explanations of the "problem" and for the "technical details". Other disciplines – such as data mining – have recently added to this body of knowledge and are considered in sections 2.4 to 2.6.

2.1. The "outlier problem"

Many researchers are faced with extreme observations in their data analysis. However, for various reasons, these observations are frequently called "outliers" in a genuine manner. Outliers affect many analysis techniques (Chase 2001). Thus, when researchers encounter extreme cases in their data analysis, they need to determine their nature, origin and how to pursue further with data analysis. In the literature, these cases are often labelled differently. Table 3 displays a list of frequently used terms.

Some authors also use different terms during the analysis process to emphasize that certain cases are still under investigation. Such terms include "suspected observations" (Iglewicz and Hoaglin 1993), "suspicious observation" (Barnett and Lewis 1994) or "potential outlier" (Hadi and Simonoff 1993).

As outliers can affect analysis techniques, some textbooks today encourage the researcher to delete these cases or average them out as disturbing phenomena (Chase 2001, Julander and Söderlund 2002) while others strongly oppose such treatment. This scientific dispute looks back on a certain history: Stigler (1973) traced the same, unresolved scientific discourse until the mid of the 19^{th} century, when researchers discussed the "correct" handling of outlying observations. Even earlier, Bernoulli (1777) questioned the assumption that errors in a dataset are identically distributed and consequently the practice pursued by his fellow researchers of discarding discordant observations without having additional information. Peirce (1852), and later Chauvenet, developed the first "objective" rejection criteria for observations based on probability calculus. Pierce, for example, proposes that…

"The principle upon which it is proposed to solve this problem is that the proposed observations should be rejected when the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection multiplied by the probability of making so many, and no more abnormal observation." (in: Beckman and Cook 1983, p124)

Term	Definition of the term
Outlier	Observation that appears to be inconsistent with the remainder of the set of data (Barnett and Lewis 1994, p7). A data point that is far outside the norm for a variable or population (Stevens 1992) A collective to refer to either a contaminant or a discordant observation (Beckman and Cook 1983). An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism (Hawkins 1980).
Extreme (observation, case, value, score)	Exhibits an extreme value, whether it is considered an outlier depends on the postulated model (Barnett and Lewis 1994)
Spurious observation	Outlier arising from inherent variability in the data (Anscombe 1960)
Contaminants	Extreme cases that are not genuine members of the population. Nevertheless, not all illegitimate cases show up as outliers and not all outliers are illegitimate (Barnett and Lewis 1994). Any observation that is not a realization from the target distribution (Beckman and Cook 1983).
Deviant case	A case that departs from a theory-based expectation and, therefore, becomes a crucial case for amendment or refutation of an established theory (Ebbinghaus 2005).
Influential observations	"Naturally occurring outliers" influencing the calculations for the "main data" Observations, individually or collectively, that excessively influence the regression equation as compared to other observations are called influential observations (Chatterjee and Hadi 1988).
Fringelier	"Unusual events that occur more often than seldom" (Wainer 1976). A special case of outliers near three standard – deviations from the mean, which is difficult to identify due to their relative proximity to the mean.
Discordant observation	Any observation that appears surprising or discrepant to the investigator (Beckman and Cook 1983).

Table 3: Terminology used for extreme cases

The critique never extinguished, and one of their main proponents can be seen in Airy (1856) who, similarly to Bernoulli, argued against the philosophy of excluding cases with "we are bound to admit all on the same terms as giving equally valid evidence." While some notion of this dispute also inspired this dissertation and is not entirely settled, criteria such as Peirce's or Chauvenet's are still taught today in many engineering programs in the United States (Ross 2003) and are in line with other researchers who pursue what Anscombe (1960) compared with a "fire insurance policy": No one expects a fire to occur in his house, yet everyone pays the premium and carries a policy just in case. The false rejection of non-outlying observations is the premium the researcher pays to protect against the possibility of sample contamination.

> *"Most modern statisticians, bemused by 5%, give rules (for the rejection of outliers) having rejection rates of about one per 20n observations. No one has explained why this should be so, no one seems to have asked" (Anscombe 1960, p127)*

Anscombe suggests that the price paid for the rejection is reasonable in light of the dangers of including extreme and spurious observations.

Overall, one can state that the scientific discourse on outliers and their "correct" handling has lasted from Bernoulli (1777) to date. Outliers have been labelled in different ways but often related to the way of "treatment" (e.g. contaminants vs. extreme observations).

Extreme observations can highly affect the inferences drawn from data as shown in Exhibit 1.

Example: Response time measurement

Reaction time analysis in social psychology often suffers from the impact of a few, very long response times. They are often thought to be generated by lapses of attention, equipment malfunction and other factor extraneous to the processes under examination in the study (Luce 1986). The use of medians or "restricted means", that means the elimination of response times outside 2 - 3 SD, is therefore a frequent practice. Because reaction time distributions are skewed and not normal, restricted means calculations result in underestimation of the population mean as only large values get eliminated by this procedure. This bias further increases with the sample size as shown in an experimental study. With large samples, observations are most likely to be excluded from the high end of the skewed distribution, whereas with small samples, observations are very unlikely to be excluded at all. A comparison of two experimental conditions with different numbers of respondents can therefore be misleading (Miller 1991).

Exhibit 1: Outliers in participant response times

The characteristics of the influence depend on the type of analysis pursued and the domain where the data are collected. More generally, the following problems can arise:

First, outliers increase error variance for the main analysis and reduce the power of statistical tests. For example, estimated regression coefficients that minimize the sum of square for error are very sensitive to outliers. Exhibit 2 outlines the problems related to underpowered studies in psychology and marketing research.

Second, if non-randomly distributed, outliers can decrease normality and violate assumptions of sphericity and multivariate normality that represent prerequisites for many statistical tests. Hence, they alter the odds of making both Type I and Type II errors (Osborne and Overbay 2004). Although according to Micceri (1989), perfect normal distributions are rare in social science datasets, they can also affect non-parametric methods. This is also underlined by Martin and Roberts (2006, p703f): "In many real-world applications, it is simply not possible to force data into a form where the three classical assumptions of linearity, homoskedasticity and normal errors are simultaneously satisfied".

Third, they can seriously influence estimates that may be of substantive interest such as correlation or regression coefficients or mean differences (see Rasmussen 1988; Schwager and Margolin 1982; Zimmerman 1994 for a more detailed discussion).

Effect sizes and statistical test power

The concepts of power and effect sizes in statistical analysis complement significance testing and have been promoted intensively by Cohen (e.g. 1988, 1992). Whereas the $p<.05$ criterion characterizes a low probability of obtaining certain results (e.g. differences in mean values) by chance, many methodologists today would argue that the reporting of effect sizes is critical as they display the magnitude and therefore the importance of the effect being tested (e.g. distance of two means or proportion of variance accounted for). Additionally, effect sizes are independent of sample size, while p-values in significance tests are vulnerable when used with large samples.

A priori calculating the power of the analysis allows determining the probability of correctly rejecting the null hypothesis. Failing to do so results in committing Type II errors or to not reject a null hypothesis when in fact it should be rejected. This may occur when sample sizes are not large enough to reliably detect expected effect sizes. Not performing these a priori calculations may therefore result in either a lack of power or a waste of resources in case of collecting too much data. A posterior, power analysis sheds light on the results: A power of .8 or .9 can raise a high level of trust in the results. For results with low power such as .2, the probability of correctly rejecting the null hypothesis is only slightly better than for falsely rejecting it. Therefore both, type I and II errors might be the consequence (Rossi 1990.; Osborne, Christiansen and Gunter 2001). Still, inadequate levels of power seem to remain present in a large number of psychological studies (Maxwell 2004).

Exhibit 2: Effect sizes and power of statistical tests

Simulation studies have demonstrated the potential impact of outliers on both parameter estimates and type I and II errors. In their study, Osborne and Overbay (2004) added 4% outliers to random samples of 50, 100 and 200 out of a large national education survey for which population parameters were known. Outliers were randomly selected from all cases which showed $| z | > 3$. The results from the correlation analyses calculated with the outliers differed heavily from those without. At a given r =.06 at the population level, the calculations for the outlier including sample ranged from r = .01 to r= -.54; at a population level r= .46, sample calculations ranged from r = .15 to r = .30. If outliers were excluded, parameter estimates were very close to the estimates for the whole population. For experimental designs and the ANOVA or T-tests, the removal of outliers produced a significant change in the T statistic primarily when a strong difference in the group means existed. In the case of almost equal means, the mean differences changed significantly. Particularly in the case of unequal group means, the calculation without outliers resulted in major reductions of type I and II errors. Similar studies have been conducted for non-parametric tests (e.g. Zimmerman 1994) and structural equation modelling (Mullen, Milne et al. 1995). The latter demonstrated, for instance, significant improvements in many of the model fit criteria by excluding three outliers at a total sample size of 98. Table 4 summarizes the effects outliers can have on statistical analysis.

Potential effects of outliers and their consequences
▪ Bias or distortion of estimates
▪ Inflation of sum of squares and hence difficulties to partition sources of variation in the data into meaningful components
▪ Distortion of p-values and statistical significance testing
▪ Faulty conclusions if unusual data have not been identified

Table 4: Possible consequences of outliers in datasets

For that reason, screening data before conducting analyses has been widely recommended and tools have been included in popular statistical software packages such as SPSS or SAS.

Extreme and outlying values sometimes, but not always, influence the results of statistical calculations. Statistical literature provides extensive knowledge on the identification, the accommodation, the rejection or the treatment of outliers (see e.g. Barnett and Lewis 1994 or Hawkins 1980 for a detailed overview).

2.2. Sources for outliers

Sources for extreme observations can be multifaceted and outlier management highly depends upon the nature of the source and whether this source can be clearly identified.

Measurement and execution errors

By the middle of the 20th century, the use of self-reporting questionnaires had reached a peak in virtually every domain of social sciences, but also other fields. Today, it remains a very popular instrument to collect data. Throughout its history, a basic concern has been the accuracy of the information acquired, or put simply: "Can we get useful information simply by asking for it?" (Willis 2005 p13). Willis summarizes the best-known sources of error which can possibly arise during a survey study as displayed in Table 5. Many of these categories relate to other means of data collection, such as observation, as well.

Major sources for error in survey data
Errors of non-observation: Are we surveying the right people? ▪ Coverage error: Certain people or units may not be included in our sample. For example, telephone surveys do not provide coverage of people who do not have telephones. ▪ Sampling error: Because we are selecting a sample, as opposed to asking everyone in the population, we may happen to selected individuals who are not representative of that population. ▪ Non-response error: If some people who are in the sample are not surveyed (because they cannot be located, refuse to participate, etc.), or if some items are left unanswered, this may introduce bias.
Errors of observation: Are we getting the right answer? ▪ Interviewer error: Interviewers may read questions incorrectly, make errors in recording data and so on. ▪ Response error: Characteristics of questions and a respondent's processing of those questions may lead to incorrect answers.
Post observation errors: Are we doing the right things with the data? ▪ Processing error: Data may be coded or analyzed incorrectly. ▪ Interpretation error: Are we drawing the correct inferences based on our obtained results? Source: Groves (1991) and Oppenheim (1966)

Table 5: Sources for error in survey data

These errors also potentially influence the occurrence of outliers. In this context, outliers are often, but not exclusively, called "contaminants", because they are considered not representing true observations. Errors can arise, as shown, from the measurement itself, but also the measurement process such as execution errors (Andrews and Pregibon 1978, Anscombe 1960, Barnett 1978): Recording errors, for example, are part of this group (e.g. a pregnancy date for a male observation), but also errors from faulty observation. In the case of a questionnaire, the design and question may be valid and reliable, but errors can occur from visual, auditory or other sensory miscomprehensions as well as more complex observation requiring the coordination of a number of senses.

Furthermore, it has been known for quite some time that peoples' responses to a measurement instrument, such as a questionnaire with rating scales, are not only related to the content, but also the instrument (e.g. Cronbach 1946, Lentz 1938). This includes phenomena such as motivated misreporting to sabotage research (Huck 2000), but also social desirability and self-presentation motives. Such motives can be powerful when reporting sensitive data. Non content-based forms of responding are usually called response styles, response sets or response biases (e.g. Baumgartner and Steenkamp 2001, Arce-Ferrer 2006). Table 6 outlines the most frequently mentioned response styles in the literature discussed by Baumgartner and Steenkamp (2001). Some of these biases are related to a particular test or to a single testing session, others depend on "reliable, stable consistencies with some generality over tests and situations" (Messick 1968). Arce-Ferrer (2006) notes that extreme response styles have also been related to the respondent's ability to adapt their own inherent systems for categorizing information to response categories as well as "collective groups filter cultural and social information in a manner that focuses on fitting with others, individuals adapt their being to match salient features of groups" (Arce-Ferrer 2006, p374f). In his cross-cultural study, Hofstede (1984) illustrates this point with the values of a specific variable, which within one society would be considered extreme or outlying cases if another society are used as point of reference. Nevertheless, the difficulty of measuring these response styles in a reliable way has been acknowledged (see e.g. Budescu, Erev and Wallsten 1997 for a detailed discussion).

Baumgartner and Steenkamp (2001) discuss ways to impede the appearance of many of these response styles such as the use of reversed scales. However, the reader is referred to their publication for a more detailed explanation.

Error-induced outliers can also arise from validity issues of conceptualization and operationalization of concepts. Studies with poorly operationalized concepts are likely to result in extreme and spurious observations. Outliers also result from sampling problems such as poor specification of the sample or collecting samples from distinctly different groups (Clark 1989). Careful examination should allow the identification of these errors and, ideally, their correction. Otherwise, treating observations as missing is preferable to deleting them. Under certain circumstances, a replacement as in Exhibit 5 can be pursued. Nevertheless, complete rejection is also frequently suggested (Clark 1989).

Looking at all the possible biases in data collection, and particularly in using questionnaires, the reader can be tempted to rather refrain from using questionnaires. Before doing so, one may note that the literature has advanced in terms of possible distortions and, therefore, “remedies” (see e.g. Willis 2005) are available. Furthermore, triangulation as one of the more recent streams in data collection techniques can also be very helpful to get a realistic picture of the phenomenon under study. “Environmental influences” on the measurement and subsequent errors have also been discussed in other data collection methods by Miller (1972) and Rosenthal and Rosnow (1969).

Overall, data collection is susceptible to errors in various stages and outliers may be the consequence. Although, triangulation, the use of recent techniques and validation in the field can help reduce these errors.

Definition of style	Theoretical explanation
(ARS) The tendency to agree with items regardless of their content. Also called agreement tendency, yea saying or positivity. (14, 21)	▪ Characteristic of stimulation –seeking extroverts who have a tendency to accept statements impulsively. (5, 18) ▪ Due to uncritical endorsement of statements by respondents who are low in cognitive abilities or have low status. (11, 18, 22) ▪ More common for items that are ambiguous, vague or neutral in desirability or for issues about which respondents are uncertain. (16, 20, 21) ▪ Most likely when respondents lack adequate cognitive resources because of distraction, time pressure and so forth.. (15)
(DARS) The tendency to disagree with items regardless of content. Also called disagreement tendency, nay – saying, or negativity. (5)	▪ Characteristic of controlled and reflective introverts who try to avoid external stimulation. (5)
(NARS) The tendency to show greater acquiescence than disacquiescence. Also called directional bias.(6, 9)	[See explanations for ARS and DARS]
(ERS) The tendency to endorse the most extreme response categories regardless of content.(7)	▪ Reflection of rigidity, intolerance of ambiguity and dogmatism.(8) ▪ Associated with higher levels of anxiety and possible deviant behaviour. (8) ▪ Characteristic of respondents with less differentiated cognitive structures and poorly developed schemas. (23) ▪ Greater for meaningful stimuli (i.e. stimuli that are important or involving to respondents). (19)
(RR) The tendency to use a narrow or wide range or response categories around the mean response. (6, 9, 29)	[Presumably similar to the explanations for ERS]
(MPR) The tendency to use the middle scale category regardless of content. (17, 22)	▪ Due to evasiveness (desire not to reveal one's true opinion), indecision (uncertainty about one's position), or indifference (lack of interest in an issue). (17, 22)
(NCR) The tendency to respond to items carelessly, randomly or non-purposefully. (13, 25)	▪ Due to lack of motivation to read the instruction and interpret items appropriately. (10) ▪ Due to partial knowledge, the frequent guesser also tends to receive higher scores on ability tests then the more cautious one. (17)
(SDB) Tendency by respondents under some conditions and mode of administration to answer in a more socially desirable direction.	▪ Unintentional distortion due to carelessness and disinterest, mood changes, changes in the depth of cognitive processing about the self or over-processing. (Dunning, Griffin, Milojkovic and Ross 1990) ▪ "Intentional impression management might involve strategically faking bad to obtain a resource"…"or faking good to make a good impression or to hide sensitive personal information". (21a, 19a)
(IJC) Interpretation of judgement categories.	▪ Preference for particular response options resulting in two people having the same viewpoint having totally different scores. (3a, 17) ▪ Differences in criticalness related to the subject to be evaluated have been discovered. (17)
(PUR) Tendency to answer purposefully in a wrong manner.	▪ Tendency to fake and distort responses in an attempt to bias the impression given to the examiner. (17) ▪ Tendency to deviate from some median or reference group average. (17)

(1) Bachman and Omalley 1984, (2) Baumgartner and Steenkamp 2001, (3) Broen and Wirt 1958, (3a) Cronbach 1950. (4) Chen, Lee and Stevenson 1995, (5) Couch and Keniston 1960, (6) Greenleaf 1992, (7) Greenleaf 1992., (8) Hamilton 1968, (9) Hui and Triandis 1985, (10) Jackson 1967, (11) Knowles and Nathan 1997, (12) Marin, Gamba and Marin 1992, (13) Marsh 1987, (14) Martin 1964, (15) McGee 1967, (16) Messick 1967, (17) Messick 1968, (18) Messick 1991, (19) O'Donovan 1965, (19a) Paulhus 1984.

Definition of style	Measurement
(ARS) The tendency to agree with items regardless of their content. Also called agreement tendency, yea saying or positivity. (14, 21)	Two general approaches: (14) ▪ Extent of agreement with many items that are heterogeneous in content (which requires that multiple scales that have little in common are available). (1, 12, 26) ▪ Extent of agreement with both positively and negatively worded items within the same scale (before negatively worded items have been reverse - scored). (26) A special case is balanced worded, logical inconsistent statement (i.e. pairs of items that are identical in substantive content, with one item worded positively and the other worded negatively) (28).
(DARS) The tendency to disagree with items regardless of content. Also called disagreement tendency, nay – saying, or negativity. (5)	Same as ARS, except that disagreement is assessed instead of agreement.
(NARS) The tendency to show greater acquiescence than disacquiescence. Also called directional bias.(6, 9)	In general, acquiescence minus disacquiescence. Most commonly measured as the mean response across many heterogeneous items. (6, 9)
(ERS) The tendency to endorse the most extreme response categories regardless of content.(7)	Number or proportion of heterogeneous items on which the respondent endorses the most extreme (positive or negative) scale categories. (1, 4, 9, 12, 24) Greenleaf (1992) suggests that the items should be uncorrelated and have equal extreme response proportions. In addition, the mean response to an item should be close to the midpoint of the scale.
(RR) The tendency to use a narrow or wide range or response categories around the mean response. (6, 9, 29)	Standard deviation of a person’s responses across many heterogeneous items. (6, 9, 29)
(MPR) The tendency to use the middle scale category regardless of content. (17, 22)	Number or proportion of heterogeneous items on which the respondent endorses the middle scale category. (4, 24)
(NCR) The tendency to respond to items carelessly, randomly or non-purposefully. (13, 25)	▪ Sum of absolute differences between responses to pairs of items, where the items in each pair are maximally correlated, have similar means across respondents, and are keyed in the same direction. 13, 25) ▪ Use (or claim to use) counting formulas penalizing guessing. (17) ▪ Direct people to answer every item except measures of personality characteristic (e.g. cautiousness). (3a)
(SDB) Tendency by respondents under some conditions and mode of administration to answer in a more socially desirable direction.	▪ Comparison of different modes of survey administration (e.g. face to face vs. self vs. computer administered). (21a) ▪ Use of additional scales or subscales of social desirability distortion (e.g. Balanced Inventory of Desirable Responding). (19a).
(IJC) Interpretation of judgement categories.	
(PUR) Tendency to answer purposefully in a wrong manner.	
(20) Paulhus 1991, (21) Ray 1983, (21a) Richman, Kiesler, Weisband and Drasgow 1999., (22) Schuman and Presser 1981, (23) Shulman 1973., (24) Stening and Everett 1984, (25) Watkins and Cheung 1995, (26) Watson 1992, (27) Wells 1963, (28) Winkler, Kanouse and Ware 1982, (29) Wyer 1969 Source: Baumgartner and Steenkamp 2001, extended with Broen and Wirt 1958 and Messick 1968)	

Table 6: Response styles

Theoretical issues

Outliers linked with theoretical issues arise from the way a set of relationships is conceptualized (Andrews and Pregibon 1978). If these do not represent the reality to a sufficient extent, or important variables are left out, outlying observations might be the consequence. The same is true for using incorrect relationships. These conditions can result in extreme observations that would not appear in a well-formulated model (Clark 1989).

Global model weakness (Beckman and Cook 1983) as one aspect of this issue deals with the inadequacy of a model to represent the phenomenon of interest, including poor development, under specification or simply wrong assumptions. In these cases, a major modification or abandoning the model is recommended (Clark 1989).

Local model weakness (Beckman and Cook 1983) arises with a generally well-specified model, but partial weaknesses. If the source of the weakness is not identifiable and correctable, the model is maintained in most cases and outliers are treated on a case-by-case basis. Examples of this kind of weakness include influential observations arising from regions in the factor space with inadequate coverage, errors in the rows of the model matrix, or large residuals. They may also reflect non-additivity or heteroscedasticity (Beckman and Cook 1983). Some authors recommend data transformation (see Exhibit 3) to a logarithmic scale to account for these outlying observations. However, this approach has to be questioned from a critical realist perspective as it changes the relationships with other variables (Atkinson 1982, 1981 and Caroll 1982 provide a more detailed discussion on outliers and transformations). Theoretical issues can be widespread and reflected in execution and measurement errors.

Data transformation in statistical analysis

Many statistical procedures assume known, mostly normal distributions. In his study on educational and psychological datasets, Micceri (1989) points out that this feature is exceedingly rare in social sciences. Consequently, researchers are subjected to commit errors even in analyses, where normality is not explicitly required (e.g. non parametric analyses in Zimmerman 1995a). Deviations can be acceptable in certain situations (see e.g. Judd, McClelland and Culhane 1995). Variable transformations are often used as remedy to solve distributional problems, even though their application raises methodological questions.

The most common transformation methods include (Osborne 2002):

- Square root transformation: First, a constant is added to bring the lowest value to +1, then values are shrinked significantly by calculating their square root.
- Log transformations: Again, a constant is added to increase the lowest value to +1 first; then the log is calculated. Higher bases for the logarithm pull in extreme values more than lower ones, frequently applied numbers include 10, 2 and e (Cleveland 1984).
- Inverse transformations: Makes small numbers very large and the other way round which reverses the distribution. Initially a variable is multiplied by -1 and a constant is added to bring the smallest value to +1 before applying the inverse transformation.

All of the methods mentioned reduce the distance between the scores. This distance is more compressed on the right side than on the left side of the distribution. If implemented correctly, the relative order of the values remains the same, nevertheless the direct interpretation becomes difficult (e.g. age, income,…) as these distances are reduced unequally.

In general, literature advises researchers to pursue data transformations with care and only to the minimum necessary to fulfil the requirements of statistical tests to be applied (e.g. High 2000). On the other hand, transformed variables change their relationship towards other variables (both transformed and non transformed ones), which can impact inferences drawn as well (see Osborne 2002 for a more detailed discussion on transformations and their use).

Exhibit 3: Transformation procedures

Natural or inherent variation in the data

Finally, an outlier may be the result of natural variation rather than weaknesses of the model (Beckman and Cook 1983). Particularly in this case, the point of reference is crucial for the evaluation of outliers. This point of reference normally consists of the remaining data and their distribution. Most distributions are expected to include outliers from natural or inherent variability within the data (Barnett and Lewis 1994). For the detection of specific alternative phenomena (e.g. high radiation values after a satellite crash in Siberia), as diagnostic indicators or special interest, the focus of the investigation may lay on the extreme observations only (Beckman and Cook 1983). While a small amount of outliers are included in known distributions (e.g. for a normal distribution about one in 150 observations will be a “mild” outlier, and only about one in 425,000 an extreme outlier), their treatment requires special attention.

Some authors even argue that most outliers actually fall in this category of extraordinary events or “influential observations” (Chase 2001). This tradition can be tracked back to

Chauvenet's simple test for outlying observations which showed that the chance of wrongly rejecting a reasonable sample value within a large sample is about 40% (Barnett and Lewis 1994, p29). Even though the assumptions of Chauvenet's test have been questioned, it still attracts attention in the literature.

> *Neyman and Scott (1971) introduced the concept of outlier-resistant distributions (i.e. distributions which are normal and uniform) and outlier-prone distributions (i.e. high probability that samples will include extremely deviant values). Once a researcher knows that a family of distributions is outlier prone, remote observations no longer appear discordant. These distributions have "heavy tails which go to zero slowly" (e.g. Neyman and Scott 1971, Green 1976 or Hawkins 1980).*

Table 7 proposes a categorization of popular distributions with respect to outlier tolerance.

Categorization of distributions related to outliers and examples	
Class 1: Absolutely outlier-resistant distributions	Normal distribution
Class 2: Relatively outlier-resistant, neither absolutely prone nor resistant	Poisson distribution
Class 3: Absolutely outlier-prone distributions, relatively resistant	Gamma distribution
Class 4: Absolutely outlier-prone, neither relatively resistant, nor relatively prone	
Class 5: Relatively outlier-prone distributions: they are also absolutely prone, but cannot be relatively resistant	Cauchy distribution
Class 6: Neither relatively resistant nor absolutely prone	
Source: Green 1976	

Table 7: Categories of distributions

Before a researcher analyzes data, knowledge about the phenomenon may reveal the kind of distribution they can expect. In case the researcher changes the type of reference distribution before analysing and reporting the data, the benchmark on which outliers are measured changes as well (Barnett and Lewis 1994). Recent literature in mathematics acknowledges that…

"Noncritical use of a hypothesis on the normality frequently results in significant mistakes, for example, in rejection of sharply divided results of observations (outlier) in statistical control of quality, and in other cases. Therefore, it is expedient to use nonparametric methods, in which to distribution functions of observations are placed rather weak requirements" (Orlov 2005, p2121).

Recent research within the data-mining literature has departed from most of the distribution assumptions as "they do not really represent reality in many cases" and advocates non parametric procedures (see section 2.4). Besides distributions, the use of certain statistical methods also implies other assumptions such as the independence of observations, which again influences the kind of distributions which can be expected. In case interdependence between the participants in a social network exists, a Pareto or power law distribution can be much more appropriate than a Gaussian normal distribution (see Exhibit 4).

Distributions in social networks

Watts (2003) shows in his study that real networks follow a "small world" phenomenon. Hereby, society is visualized as consisting of weakly connected clusters, each having highly interconnected members. This structure allows cohesiveness and speed/spread of information across the whole network. In their initial model, Watts and Strogatz also assume that links are Gaussian distributed. Barabási, Albert, Jeong and Bianconi (2000), however, studying the World Wide Web, find that the structure shows a power law distribution, where most nodes have only a few links and a tiny minority – the hubs – are disproportionately very highly connected. The system is called scale-free; no node can be taken to represent the scale of the system. Defined as a "scale - free network", the distribution shows (nearly) infinite variance and the absence of a stable mean. It turns out that most real life small world networks are scale free (Ball 2004) and fractal (Song, Shlomo and Makse 2005). Those networks include phenomena in the field of epidemiology, metabolism of cells, internet and networks of sexual contacts (Barabási and Bonabeau 2003).

Exhibit 4: Non Gaussian distribution example in social research

The size of the sample drawn also impacts the probability of recording a "true" outlier. In the case in which a researcher draws a wider net and the dataset increases, the likelihood of outlying values will also increase. For instance, 1% of the cases on a normally distributed scale lies outside of 3 standard deviations, the more cases the researcher collects, the higher the probability of catching one of the extreme ones (Osborne and Overbay 2004).

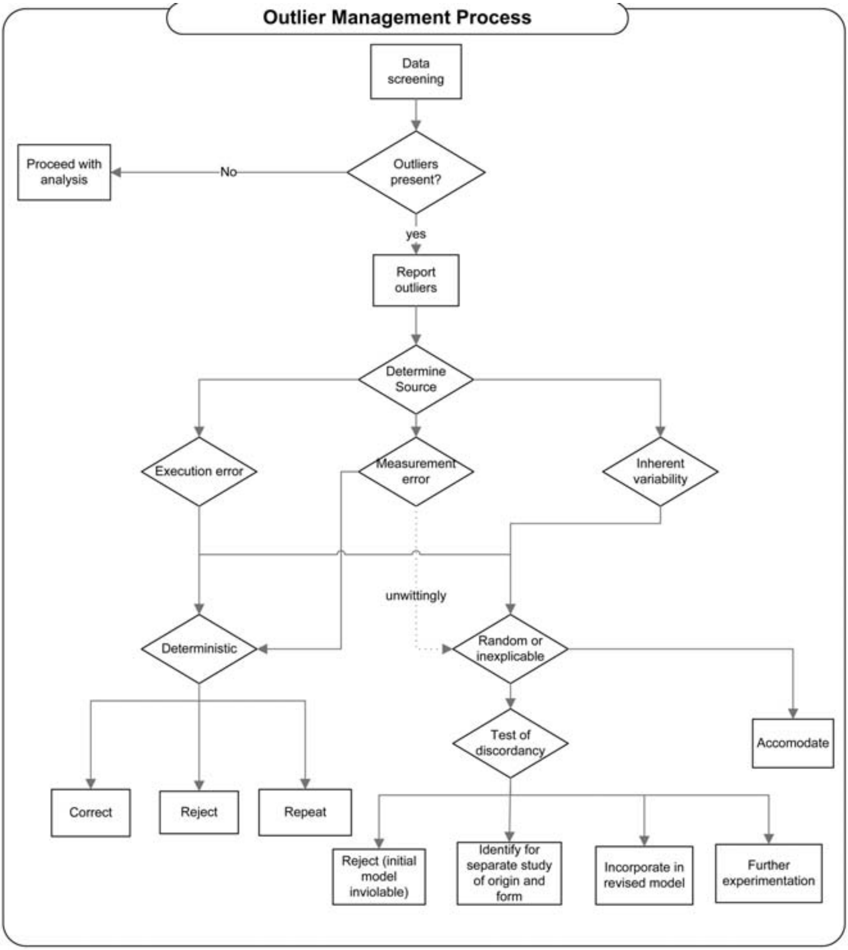
To evaluate whether outliers arise from misspecification, data collection error or inherent variability, the identification of the source of variation is key. Assumed distributions allow to compare whether observations may be suspicious, but do not reveal the source of variations.

As sources for outliers heavily influence the whole outlier management process, they require primary attention. The next section presents an outlier management process consistent with the recommendations in statistical literature.

2.3. Outlier identification and management in statistics literature

The main objective of presenting the different topics on outlier management in a sequential manner is to give the reader a more “hands on” understanding of the interaction of the different steps. Similar approaches can be found in Clark (1989). The literature agrees that all samples should be subjected to some type of outlier procedure before the main analysis (Beckman and Cook 1983), even though some sources admit that common practice is to mechanically exclude cases exhibiting extreme characteristics such as values more distant than three standard deviations from the mean (Lin and Brown 2006). When multiple outliers are suspected, median and MAD scale are frequently suggested (Andrews, Bickel, Hampel, Huber, Rogers and Tukey 1972 in: Lin and Brown 2006). The process will be treated in the subsections that follow, explaining each step individually.

Figure 1 illustrates a systematic approach to outlier management adapted from Barnett and Lewis (1994).



Source: Adapted from Barnett and Lewis (1994)

Figure 1: Outlier management process

Identification

For many data analysis ventures, the first step consists of screening the data through visual inspection or exploratory data analysis. Depending on the main analysis to be pursued later, a related outlier test is recommended here. For instance, what seems to be an outlier under a univariate perspective can seem completely normal in a multivariate view.

A broad body of literature has discussed the identification of outliers, including visual inspection and various forms of discordancy tests (Barnett and Lewis 1994, chapter 4), i.e. procedures providing information on whether a suspected observation is indeed different from the remainder of the data and can therefore be regarded as outlying or not. Most statistical identification tests are built around an evaluation of the separation of the observations from the sample and of the spread of the sample, as well as the influence of a particular observation in an analysis. This makes intuitive sense as the nature of outliers and extreme cases is relative and arises only in relationship with other values or by comparing them to a given distribution. The formulation of specific test statistics depends on the distribution of the sample data, the type of analysis to be performed, and the properties and distributions of the test statistic in a given situation. The construction of the outlier identification test statistic revolves around a consideration of the sample excess (the difference between the outlier and its nearest neighbour), range, variance, sums of squares, skewness and kurtosis (Tietjen and Moore 1972, Barnett and Lewis 1994). These tests are also often distinguished into the “mean shift models” group (in particular for linear models) or “variance-inflation model” group (Beckman and Cook 1983). Table 8 displays some frequently applied methods for outlier identification that can help the researcher to focus on candidate observations.

Type of analysis	Procedure	Description
Univariate, metric	Outside +/- 3SD	Values outside three SD are screened for “correctness” + Easy to apply, good initial screening tool (Osborne and Overbay 2004). - May cause problems with certain distributions and small samples (Van Selst and Jolicoeur 1994)
Multivariate, metric	Influence or distance statistics	Mahalanobis’ distance, Cook’s D Calculates the influence each case has on analysis (Newton and Rudestam 1999).
Univariate and Multivariate	Standardized residuals	For regression analysis, $z > 3$ is frequently used; For ANOVA type analyses, within cell outliers (distance from subgroup) identified through residuals (Tabachnick and Findell 2000).

Table 8: Frequently used outlier identification procedures

However, regardless of the formulation of the statistic, outlier evaluation remains subjective (Collett and Lewis 1976, Finney 2006) and depends substantially on our understanding of the problem (Beckman and Cook 1983). Therefore, knowledge about the phenomenon to be analyzed is very important to evaluate the existence of outliers and “allow a distinction on whether outliers are assumed to carry some information about the parameters in the basic model or not” (Beckman and Cook 1983, p123). Statistical procedures can assist this identification process, especially if several variables are involved but never replace knowledge about the phenomenon and intuitive judgment. Table 9 and selected publications (e.g. Barnett and Lewis 1994 chapter 4 or Hawkins 1980) give a more detailed overview on identification procedures.

Outlier identification literature for selected procedures					
Univariate Data	Multivariate Data	Regression/Anova	Contingency tables	Time Series Data	Other Procedures
Normal Distributions: McMillan and David 1971, Tietjen and Moore 1972, Moran and Mcmillan 1973, Rosner 1975, Rosner 1983, Hawkins 1980; Barnett and Lewis 1994, Bendre and Kale 1987	*Normal Distributions:* Barnett 1979, Hawkins 1980, Schwager and Margolin 1982, Barnett 1983b, Bacon-Shone and Fung 1987, Rousseuw and Zomeren 1987	*Normal Distributions*: Brown 1975; Schweder 1976; Andrews and Pregibon 1978; Carol 1980; Hawkins 1980; Chambers and Heathcote 1981; Butler 1983, Marasinghe 1985; Jennings 1986; Rousseuw and Zomeren 1987; Barnett and Lewis 1994, Dupuis and Hamilton 2000, Schwertman, Owens and Adnan 2004	Haberman 1973; Brown 1974; Gokhale and Kullback 1978; Fuchs and Kenett 1980; Kotze and Hawkins 1984; Bradu and Hawkins 1985; Mosteller and Parunak 1985; Simonoff 1988	Fox 1972, Chernick, Downing and Pike 1982; Muirhead 1986; Barnett and Lewis 1994, Tsay 1986; Chang, Tiao and Chen 1988, Tsay, Pena and Pankratz 2000, Chen and Liu 1993, Tsay 1988, Chan and Cheung 1994	*Structural equation modelling*: Kline 2005 p 51, Bollen 1990, Mullen, Milne et al. 1995. *Bayesian Models*: Box and Tiao 1968; Abraham and Box 1978; Freeman 1980, Pettitt 1988 *Cluster Analysis:* Gnanadesikan 1977., Milligan and Hirtle 2003.
Non-normal Distributions Kabe 1970, Dornbos 1976, Barnett and Lewis 1994, Hoaglin, Iglewicz and Tukey 1986	*Non – normal Distributions* Barnett 1983b, Barnett 1983a, Barnett and Lewis 1994, Sinha 1984				
Source: adapted and completed from Clark 1989					

Table 9: Outlier identification in literature

Outlier identification procedures can be affected by “masking” and “swamping” problems. “Masking” means that one very extreme outlier that is first detected by a test procedure may hide another, less extreme one (e.g. Hoaglin, Iglewicz et al. 1986). If the researcher attempts to analyse two outliers together, attention needs to be paid to “swamping” (Fieller 1976), i.e. when one of two observations is outlying, but the other not. The more extreme one may influence the mean of the two so heavily that it “swamps” or carries the non-outlying value with it and the researcher concludes that both observations are too extreme. Table 10 summarizes literature on these issues.

Even though sophisticated procedures for various distributions and analysis situations have been developed, the final decision on whether a case is considered outlying remains with the researcher.

Outlier masking literature for selected procedures					
Univariate Data	Multivariate Data	Regression/Anova	Contingency tables	Time Series Data	Other Procedures
Tietjen and Moore 1972, Rosner 1975; Rosner 1983; Bendre and Kale 1985; Bendre and Kale 1987, Bacon-Shone and Fung 1987	Bacon-Shone and Fung 1987	Atkinson 1986; Hoaglin, Iglewicz et al. 1986	Simonoff 1988	Chen and Liu 1993	
Source: adapted and completed from Clark 1989					

Table 10: Outlier masking literature

Independent of a specific methodological perspective, the researcher only *proceeds with the analysis* in the event that no outliers have been identified. Otherwise, *reporting outliers* is highly recommended:

> *"As to practice, I suggest that it is of great importance to preach the doctrine that apparent outliers should always be reported, even when one feels that their causes are known or when one reject them for whatever good rule or reason" (Barnett and Lewis 1994, p40)*

Outlier management

The next steps in the process refer to outlier management procedures: The literature often distinguishes between accommodation, incorporation and rejection. In a first step, the literature suggests the identification of *the source of the outlying case*, as the choice of outlier management highly depends on it (see e.g. Grubbs 1969). In case the researcher cannot determine the source of the outlying value, he may encounter the following or a similar dilemma:

If the outliers had proved to be discordant on an assumed normal distribution it is quite likely (bearing in mind the possibilities of inexplicable 'gross errors') that we would have chosen to reject them before proceeding to further study of the data. We cannot, of course, be sure that this action is entirely proper (Barnett and Lewis 1994, p39).

If an execution or measurement error can be determined, the researcher may either *correct* the problem (in case of e.g. an erroneous coding where the original data record is still available), partially or completely *repeat* the measurement (e.g. re-interview the respondents relevant to the extreme cases) or *reject* the respective cases. While the first two options do not impact the research approach, rejection results in an incomplete dataset and the random sample is turned into a censored one, which can produce serious methodological problems. Some very early authors in the field recommended sample trimming (i.e. the exclusion of a fixed percentage of lower and upper extreme sample values) before even starting with the main analysis. A correction by interpolation from the remaining dataset (such as the replacement with the nearest neighbour, a statistical equivalent or the use of a replacement procedure for missing values), the dataset at hand must be questioned from a strict methodological point of view. It modifies original data and consequently also relationships to other variables.

In the case that execution or measurement errors cannot be determined and the source for outliers is considered *random* or *inexplicable*, the literature either suggests an *accommodation approach* or *discordancy testing*.

Accommodating outliers implies reducing their effect without discarding them if they influence the analysis so heavily that distortion occurs. For many, especially older, statistical procedures, departures from assumptions such as normality or outliers can be very misleading. Therefore, a careful consideration of more recent methods is recommended (Wilcox 1998). Non-parametric methods, for example, are robust and often a good solution, but they are less known and applied. Accommodation procedures are often based on some sort of "reweighting", i.e. weighting outlying observations iteratively with a coefficient of less than one but greater than zero. They are also referred to as "estimators based on residuals and moments" (Beckman and Cook 1983). "Winsorizing" – i.e. the replacement of an outlying observation by the value of its nearest neighbour or "methods based on order statistics" (Beckman and Cook 1983) – are frequently mentioned as a second group of procedures. As a result, they are robust or relatively unaffected by outliers and, therefore, often called robust procedures. "Robustness signifies insensitivity to small deviations from the assumptions. Primarily, we are concerned with distributional robustness: the shape of the true underlying distribution deviates slightly from the assumed model (usually the Gaussian law)" (Huber 2004, p2). As a result, the effects of the extreme observations are changed. Nevertheless, robust procedures extend beyond the treatment of outliers. Other popular means of accommodation include the use of mixture models and mean estimates to provide some protection against outliers. In comparison to simple identification, accommodation requires that much information about the process generating the outliers is known before it is executed (Beckman and Cook 1983). Further options can include the transformation through square roots and logarithms in order to soften the impact of outliers. The transformation of a variable not only makes a distribution less skewed, it also changes the relationship between the original variable and the other variables in the model (High 2000). Table 11 displays outlier accommodation literature.

Outlier accommodation literature for selected procedures					
Univariate Data	Multivariate Data	Regression/ Anova	Contingency tables	Time Series Data	Other Procedures
Normal Distributions: Anscombe and Barron 1966, Guttman 1973b, Barnett and Lewis 1994, Bendre and Kale 1987 Non Normal Distributions Joshi 1972; Kimber 1983; Barnett and Lewis 1994	*Normal Distributions:*Gnanades.R and Kettenri.Jr 1972, Guttman 1973b; Barnett 1979; Cleveland 1979; Rousseuw and Zomeren 1987; Barnett and Lewis 1994, *Non – Normal Distributions:* Barnett 1979, Barnett 1983a	Schweder 1976; John 1978, West 1984, Barnett and Lewis 1994, Mahajan, Sharma et al. 1984, Giltinan, Carroll and Ruppert 1986; Rousseuw and Zomeren 1987	Emerson and Hoaglin 1983, Mosteller and Parunak 1985	Denby and Martin 1979, Barnett and Lewis 1994, Basawa, Huggins and Staudte 1985, Schmidt 1986, Tsay 1986, Rousseuw and Zomeren 1987, Chang, Tiao et al. 1988, Chan and Cheung 1994, Dijk, Franses and Lucas 1999	*Robust SEM*: Huba and Harlow 1987 *Robust Bayes Time series:* Carter and Kohn 1996, Chow 2004
Source: adapted and completed from Clark 1989					

Table 11: Outlier accommodation literature

<u>Discordancy tests</u> are frequently set up by using maximum likelihood and the principle of local optimality. Recent literature proposes sophisticated bootstrap procedures which take "into account pattern in the underlying data and so the effect of skewed or heavy-tailed error distributions in modelling should be less likely to produce a larger number of spurious outliers

than under traditional approaches" (Martin and Roberts 2006, p703f.). A test of discordancy can provide valuable information for the decision how to proceed with outlier treatment, e.g. the likeliness whether an extreme observation belongs to a population or not. Still, such a test only gives a statistical, but not a theoretical argument.

After discordancy testing, the following alternatives can be relevant for outlier treatment: (a) the subsequent rejection, (b) important new information contained in accompanying variables that would otherwise have gone unnoticed and, therefore, require a separate study, (c) its incorporation through a revision of the model (global weaknesses) or method of estimation, or (d) a recognition of an inherent weakness in the data and, thus, to further experimentation (Beckman and Cook 1983). These options also apply for outlying observations arising from inherent variability.

If outliers are accepted as part of the phenomenon, the researcher decides to incorporate them. This means that the respective calculations are realized including extreme and non-extreme cases. "The current model under which the outliers were produced is rejected in favour of a better model that accounts for the extreme observations. The outliers are left as they are, but the context is changed so that they are no longer perceived as extreme. Incorporation is the recommended procedure when global model weakness is suspected as the cause of the extreme observations" (Clark 1989) Incorporation can be seen related with discordancy testing. The working hypothesis (or null hypothesis) of the discordancy test will express some basic probability model for the generation of all the data with no contemplation of outliers, the alternative hypotheses express ways in which the model may be modified to incorporate or explain outliers as reflections of contamination (Barnett and Lewis 1994, p43). For the alternative hypotheses, the literature proposes a wide array of possibilities, which depend a lot on the type of analysis and theoretical background. For a more detailed overview on criteria and principles of discordancy tests, the reader is referred to Barnett and Lewis (1994, chapter 6). Several alternatives can be useful here:

The *mixture alternative* reflects the option that the sample may include an unexpected degree or form of inherent variability not represented by the basic model, i.e. that observations arise from another distribution to that assumed for the basic model (refer to Box and Tiao 1968 and Guttman 1973a for a more detailed discussion). The *slippage alternative* is the most common alternative used in publications: "In its most usual form the slippage alternative states that all observations apart from some prescribed small number k (1 or 2) arise independently from the initial model *F* indexed by location and scale parameters, μ and σ^2, whilst the remaining *k* are independent observations from a modified version of *F* in which μ or σ^2 have been shifted in value (μ in either direction, σ^2 typically increased)" (Barnett and Lewis 1994, p49). The *exchangeable alternative* can be seen as an extension of the slippage alternative and is concerned with estimating mean and variance in an exponential distribution in an outlier robust manner. (Barnett and Lewis 1994, p51). Other, less popular, alternatives are summarized by Barnett and Lewis (1994, p52).

Rejecting outliers can be seen as a re-weighting procedure in which the outliers are assigned a zero coefficient. This results in discarding the outlier from the subsequent analysis and is recommended for outliers arising from errors or sample contamination if the retention would likely result in invalid inferences (Clark 1989). Such procedures frequently include – for example – trimming (i.e. the exclusion of a certain percentage of extreme observation of a sample). Treating outliers as missing values is similar and a common practice. Table 12 details more literature for this handling decision.

Outlier rejection literature for selected procedures					
Univariate Data	Multivariate Data	Regression/ Anova	Contingency tables	Time Series Data	Other Procedures
King 1953, Proschan 1953, Anscombe 1960; Ferguson 1961a; Ferguson 1961b, Dixon 1962, Rousseuw and Zomeren 1987	Anscombe 1960; Ferguson 1961a; Ferguson 1961b,Rousseuw and Zomeren 1987	Anscombe 1960; Ferguson 1961a; Ferguson 1961b,Stefansky 1972; Mirvalieu 1978, Galpin and Hawkins 1981	Mosteller and Parunak 1985	Pena and Maravall 1991	Structural equation modelling: Kline 2005
Source: adapted and completed from Clark 1989					

Table 12: Outlier rejection literature

"The outright rejection of outliers has statistical consequences for the further analysis of the reduced sample. We would no longer have a random sample, but a censored one. The practice of replacing rejected not deterministically explicable outliers by statistical equivalents (further simulated random observations from the assumed underlying distribution) involves similar consequences" (Barnett and Lewis 1994, p40).

<u>Treating outliers as missing values</u>

Some outliers are caused by – for example – recording errors and cannot be corrected using information from the data collection phase (e.g. by re-entering data from a paper and pencil questionnaire). Treating these cases as missing values is common practice; nevertheless, many statistical tests are very sensitive to missing values, especially if the sample size is already limited. Therefore, researchers frequently seek to replace these missing values through information existing in the rest of the cases. Various techniques have been developed to extrapolate these values from the remaining dataset; they are also often called imputation.

Before imputation is started, missing values are investigated as to whether their missing status exhibits patterns or can be considered random. In the second case, replacement by imputation is recommended. This can result from either values "missing completely at random" (MCAR) or "missing at random" (MAR) In the first case (MCAR), a "missing" value does not depend on the variable itself or on the values of other variables in the database (Little and Rubin 1987) In the second case (MAR), the probability of missing data on any variable is not related to its particular value and the pattern of missing data are traceable or predictable from other variables in the database. Hence, multiple imputations are used to fit the missing data on the rest of the data to be modelled. Depending on the goal of the analysis, different techniques are used. Lemieux and McAlister (2005) used the most prominent techniques in a simulation study that resulted in the following recommendations: If the objective is to get the most accurate imputations, or imputations yielding the most accurate estimates of means or co-variances, then expectation maximization (EM) should be used. To get the most accurate estimates of variances, then hot deck (HD) is more appropriate. To obtain the most accurate model coefficients or models producing the most accurate model predictions, then Mean should be used. Completely dropping cases for which information is missing resulted in the worst performance of all techniques under investigation.

Exhibit 5: Treating outliers as missing values

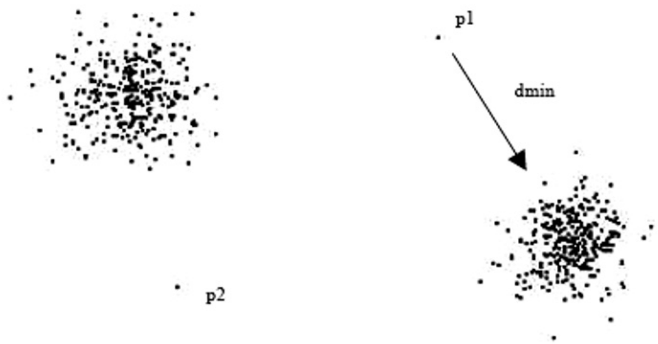
In case none of the options mentioned before are applicable due to – for example – methodological or theoretical constraints, the researcher may consider *further experimentation*, i.e. reworking the framework for data collection and re-collecting data.

Altogether, statistics literature has developed a broad array of technical knowledge on the phenomenon of extreme cases and outliers that can assist the researcher in determining the nature of the extreme value. Nevertheless, the final decision on how to deal with the outlying observation is left to the researcher and, therefore, subjective.

2.4. Outliers in data-mining literature

The data mining and computing community has built on the knowledge developed by the statistics community. The focus of this community centres on the applications of multivariate distributions, whereas the statistics community emphasized the univariate- and distribution-based perspectives to a larger extent. This limitation was criticized by the data-mining community (Knorr and Ng 1998) that recognized the mostly non-parametric nature of data (See for example Knorr and Ng 1999). Researchers linked outlier and extreme value identification with clustering procedures. First, they introduced the concept of distance -

based outliers (see Figure 2): An object p in a dataset D is a distance - based outlier if at least a percentage pct of the objects in D lays greater than distance dmin from p.



Source: HS KDD group LMU Munich 2002

Figure 2: Distance - based outliers

Knorr and Ng (1999) proposed using an example's distance to its nearest neighbours as a measure of unusualness (e.g. Angiulli and Pizzuti 2002, Eskin, Arnold, Prera, Portnoy and Stolfo 2002, or Knorr, Ng and Tucakov 2000). From this concept, they derived global outliers (related to the whole amount of data) and local outliers (related to data in the neighbourhood) (Breunig, Kriegel, Ng and Sander 2000).

Furthermore, they added "depth based outliers" to the existing terminology. Based on some definition of depth (see e.g. Tukey 1977), data objects are organized in layers in the data space, with the expectation that shallow layers are more likely to contain outlying objects than the deep layers. For example, objects which are nearer to other cases have higher depths than those much further away. Nevertheless, efficient algorithms focus on two or three dimensions and become inefficient for four and more dimensions as for the complex calculations of multi-dimensional convex hulls (Breunig, Kriegel et al. 2000).

Deviation based techniques identify outliers by inspecting the characteristics of objects and consider an object that deviates from these features as an outlier (Arning, Agrawal and Raghavan 1996).

Support vector based procedures have been introduced as a reaction to the need of treating a large dataset. Here, a set of vectors defines the borders for "normal" data points (Tax and Duin 1999). Table 13 summarises the streams just mentioned as well as other research directions arising from data-mining literature. These categories may overlap to some extent (adapted from He, Xu and Deng 2005).

Approach	Description
Distribution based approaches	Deployment of some standard distribution (e.g. normal) and labelling deviating points. Mostly found in the statistics literature, but also in combination with supervised learning approaches in data mining literature (e.g. Hawkins 1980; Barnett and Lewis 1994 Rousseuw and Zomeren 1987; Yamanishi, Takeuchi and Williams 2000; Yamanishi and Takeuchi 2001).
Depth based approach	Data objects are organized in convex hull layers in data space according to peeling depth, and outliers are identified as data objects with shallow depth values (Nuts and Rousseuw 1996; Johnson, Kwok and Ng 1998).
Deviation based techniques	Outliers are identified by inspecting the characteristics of objects and consider an object that deviates from theses features as an outlier (Arning, Agrawal et al. 1996).
Distance - based method	Outliers are determined based on the distance from the k^{th} nearest neighbour or from the sum of distances from the k^{th} nearest neighbour (Knorr and Ng 1997; Knorr and Ng 1998; Knorr and Ng 1999; Knorr, Ng et al. 2000, Ramaswamy, Rastogi and Kyuseok 2000; Angiulli and Pizzuti 2002; Bay and Schwabacher 2003).
Density based outliers	A local outlier factor is derived for each case, depending on the local density of its neighbourhood (Breunig, Kriegel et al. 2000; Jin, Tung and Han 2001; Tang, Chen, Fu and Cheung 2002; Hu and Sung 2003; Papadimitriou and Faloutsos 2003).
Clustering based outliers	These techniques regard clusters as small outliers or identify outliers by removing clusters from the original dataset. Cluster based outliers are related to this concept (Jiang, Tseng and Su 2001; Yu, Sheikholeslami and Zhang 2002; He, Xu and Deng 2003).
Sub space based	Identifying outliers by observing the density distributions of projections from the data e.g. by utilizing frequently appearing patterns to identify outliers in high dimensional space or by using a hypergraph model to detect outliers in categorical data (Aggarwal and Yu 2001; Wei, Qian, Zhou, Jin and Yu 2003; He, Xu, Huang and Deng 2004).
Support vector based	Support vector novelty detector (SVND) was recently developed and proposed by Tax and Duin (1999; Scholkopf, Platt, Shawe-Taylor, Smola and Williamson 2001; Cao, Lee and Chong 2003). This vector defines the border of "normal objects" and therefore allows a distinction to outliers.
Neural Network based	The replicator neural network (RNN) is employed to detect outliers (Harkins, He, Williams and Baster 2002; Williams, Baster, He, Harkins and Gu 2002).

Table 13: Research streams on outliers in the data-mining literature

Data-mining literature as a growing, but relatively young domain has used the investigation of outliers to discover - for example - credit card fraud, irregularities in large datasets for e-commerce transactions or network intrusion (Aggarwal and Yu 2001, Bolton and Hand 2002, He, Xu et al. 2003, Eskin, Arnold et al. 2002), but also anomalies in medical datasets (Liu, Cheng and Wu 2002) and criminal incidents (Lin and Brown 2006).

> *"Most studies in KDD (Knowledge discovery in databases) focus on finding patterns applicable to a considerable portion of objects in a dataset. However, for applications such as detecting criminal activities of various kinds (e.g. electronic commerce), rare events, deviations from the majority, exceptional cases may be more interesting and useful than the common cases. Finding such exceptions and outliers, however, has not yet received much attention in the KDD community as some other topics have, e.g. association rules" (Breunig, Kriegel et al. 2000, p93).*

2.5. Outliers in economics literature

The development of the economics literature on outliers is very closely linked to statistical procedures, such as regression and time series analysis. Similar to the statistics community, the question is raised as to how large values are dealt with and if they affect, for example, parameter estimates. Seasonality, major events (such as economic crises, wars or strikes) as well as gross errors in measurement and the collection and processing of the data are of high interest to the economic research community. Nevertheless, they have the same impact on the results as standard methodology for time series analysis (Battaglia and Orfei 2005). In time series analysis, outliers can cause more problems for detection as successive observations are correlated. Beginning with Fox (1972) and Tsay (1988), many authors (e.g. Chen and Liu 1993, Chan 1995, Battaglia and Orfei 2005) distinguish between additive outliers, innovative outliers, level shifts and temporal change in time series data. An additive outlier affects only the level of a particular observation but subsequent observations in the time series are not affected. An innovation outlier affects all observations beyond a certain time point through the memory of the underlying process (Maddala 1998, p426). "A level shift consists of an event in which the effect becomes permanent on time series values, and a temporary change is an event having an initial impact, the effect of which decreases exponentially according to a fixed dampening parameter" (Battaglia and Orfei 2005, p107f).

2.6. Outliers in social science literature

Outliers have been a concern for some researchers involved in developing a sound methodology for empirical tests, particularly in the field of psychology. Parametric tests and ordinary least square regression analysis have dominated the literature, even though their problems with deviations from the assumptions are well known (Wilcox 1998). Empirical reviews criticize that only about 8% of the publications check the assumptions required by the test procedures in use (Osborne, Christiansen et al. 2001). This is even more serious in the light of empirical evidence that datasets in psychological and educational journals rarely follow normal distributions. “Tail weights from the uniform to the double exponential, exponential-level asymmetry, severe digit preferences, multimodalities and modes external to the mean/median interval” were identified as forms of deviations (Micceri 1989, p155). Meanwhile, modern and “less problematic” methods such as trimmed sample or rank based methods have been developed, but the uncritical use of parametric tests is sustained (Wilcox 1998, Zimmerman 1994).

While opinions converge that outliers can be problematic for standard analysis procedures if present in a dataset and not handled properly, researchers disagree on the way of identification and handling. Some argue that simple procedures such as Mahalanobis’ distance or Cook’s D (Newton and Rudestam 1999) or “$| z | \geq 3$” can be quite effective for outlier identification (e.g. Osborne and Overbay 2004), others demonstrate insufficiencies in this area (e.g. with highly skewed distributions by Miller 1991) and develop alternative ways such as cut-off points (Van Selst and Jolicoeur 1994).

In terms of outlier management, some studies report the usefulness of deleting outliers: In military personnel test data, Lance, Stewart and Carretta (1996), for instance, did not identify any impacts on the predictive capability of the model in use and therefore recommend that outlier treatment can be ignored if samples are large enough and the tests are reliable enough to be used in this area. Others find that even if outliers are legitimate or their cause is unclear, their removal may be the “most honest estimate of population parameters”(Judd and McClelland 1989 in Osborne and Overbay 2004). This bias in opinions is also reflected in one article from the field of psychology: Orr, Sackett and Dubois (1991) administered a survey to authors in two well-respected psychology journals. First, the authors were asked which of the three options would come closer to their belief in case they would encounter outliers during their data analysis. Table 14 displays the result for this question.

Attitudes towards data removal	
29%	All data points always should be included in an analysis regardless of where they lie relative to other data points.
67%	Data points should be removed if they are extreme outliers and there is an identifiable reason that leads you to consider them invalid.
4%	Data should be removed from an analysis if they lay in an extreme area relative to the rest of the data. There does not need to be identifiable reason to believe that they are invalid. Extremity is reason enough.
N = 100; Questionnaire submitted to senior authors of all published papers using correlation or regression type analysis in the Journal of Applied Psychology and Personnel Psychology from 1984 to 1987; Source: Orr, Sackett et al. 1991	

Table 14: Authors' attitudes towards outlier handling

Afterwards, respondents were asked to rank those techniques they use in terms and order from a list of 10 outlier detection techniques. Techniques not used were not to be ranked. Table 15 displays the results.

Ranking of techniques for assessing outliers in bivariate relationships					
Technique	Ranked #1	Ranked #2	Ranked #3	Ranked # 4	Total
Scatter plots	70	2	5	4	81
Plots: residuals against predicted values	7	38	6	4	55
Plots: residuals against independent variables	3	11	21	8	43
Standardized Residuals	1	8	20	16	35
Mahalanobis' distance	0	2	2	10	14
Studentized residuals	1	2	1	8	12
Deleted residuals	0	1	0	8	9
Cook's D	0	1	1	5	7
Campbell's Q	0	0	1	4	5
Leverage values	0	0	0	3	3
Source: Orr, Sackett et al. 1991					

Table 15: Authors' attitudes towards outlier handling (2)

Authors who declared "all data points should always be included" also answered the second question. Therefore Orr, Sackett et al. (1991) speculated that "a number of respondents may have made rankings based on familiarity rather than actual use" (p473f). In a second study, the authors unveiled that the removal of outliers can influence effect sizes measures in individual studies.

In terms of outlier management, procedures similar to those in the statistics literature are presented:

Zimmermann (1995a; 1995b), for example, investigated non-parametric procedures and found them to be affected by outliers as well, even though non-parametric procedures remain more powerful than, for instance, F or t-test if distributions are heavy tailed. As a consequence, he proposed down-weighting outliers. He nevertheless concludes that if data were generally transformed to ranks, parametric tests would lead to the same results.

However, an article by Wilkinson appeared in 1999 and aimed to introduce the new statistical sections of the American Psychological Association publication manual. Therein, the concern for outliers was laid down in the sentence:

> *"The use of techniques to ensure that the reported results are not produced by anomalies in the data (e.g., outliers, points of high influence, non-random missing data, selection bias, attrition problems) should be standard component of all analyses." (Wilkinson 1999, p599)*

However, the article remains rather unspecific on the procedures to be applied except for a few lines later, where the author emphasizes the importance of data screening and underlines the importance of not using "this opportunity to discard data or change values". This last paragraph can be interpreted in the sense that any data modification, discarding, or change in a larger sense must be reported. In a recent version of the handbook of psychology, three alternatives are given for the treatment of outliers: "One can eliminate those elements that appear to be outliers to the overall set of data. Alternatively, the relationship between the obtained clusters and the suspected outliers can be investigated after an initial clustering is completed. A third alternative is using a clustering method resistant to the presence of outlier" (Milligan and Hirtle 2003, p175).

In other areas of social science, outliers have seen much less attention: One comment in the American Sociological Review (Kahn and Udry 1986) explicitly criticizes another researcher's findings for being contradictory with existing knowledge and points the reasons towards not having pursued a proper outlier check before analysing the data. In a reply, this argument is refused by the author of the original manuscript (Jasso 1986).

A recent publication in Political Research Quarterly by Lawrence (2007) critically investigates a previous model (Rothenberg and Sanders 2000) that gave substantial evidence that lame duck republican members of Congress engaged in shirking in the impeachment process of former president Bill Clinton. By using a different estimation method or excluding one single outlier – the unusual behaviour of one member of congress – the result was reversed. Another application can be found in Ebbinghaus's (2005b) cross-national comparison, which suggests a focus on the exceptional (outlying) case through comparative, case-based analysis. The focus on the individual case allows for the accounting of the understanding of mechanisms and the complexity of social reality.

Overall, one notes that the attitudes towards outliers and their handling vary in social science literature. Clear-cut advice seems difficult to find and identify, the APA manual only gives tendencies.

2.7. Outliers in business-related literature

Literature in business studies has recognized the outlier phenomenon to a very limited extent. The best-known studies come from areas closely linked to economics and statistics, such as finance and business forecasting.

Finance literature has recognized the importance of inspecting deviating values. Standardized unexpected earnings (SUE) studies of common stock performance have used outlier identification techniques for that reason (Bidwell 1979, Lantané and Jones 1977, Lantané and Jones 1979). Standard unexpected earning measures have been found useful in predicting abnormal stock returns. Many other publications on SUE, such as Bird, McElwee and McKinnon (2000) demonstrate how to identify investment opportunities based on earnings surprises and highlight the extent to which the opportunities differ across countries.

Business forecasting unveiled the value of naturally occurring outliers and influential observations and recommended that they should not be discarded. Chase (2001) argued that they might be excellent resources for understanding the dynamics of the marketplace, especially in data-rich environments.

Sharma and Mahajan (1980) developed a method to derive early warning indicators for future business failures from corporate performance indicators, such as return on assets and current ratio. From a larger number of performance variables, values above a certain level for these two variables allowed for the successful prediction (with a 73.9% - 91.7% probability) of whether a retailing company would fail within five years.

In marketing, Allenby and Ginter (1995) used extremes to design products and to segment markets. Studying deviant cases from a regression line, Julander and Söderlund (2002) found that for about one third of the sample, trust could serve as glue for satisfaction even after very bad experiences in a service encounter. This study stands out for seeking theoretical contributions through the study of outliers.

Mullen, Milne et al.'s (1995) article deals with outlying values and proposes outlier identification and management procedures for structural equation modelling. They argue somewhat in favour of cleaning of a dataset from influential observations. Julander and Söderlund's (2002) study takes the opposite path and argues in favour of using those outlying and influential cases to improve their model on customer satisfaction.

Overall, business related literature has paid very limited attention to the outlier phenomenon.

3. Summary and implications

The section on extreme phenomena showed that research in marketing and other domains has demonstrated a substantial interest in people and observations that behave highly different to the mainstream. Thus, surveying the mainstream and capturing one of these observations would consequently result in an observation that is very different – and probably outlying – to the others. Section 2.2 overviewed the relevance and handling of extreme data points in different areas of research. Statistics and data mining literature have built up a very extensive body of knowledge on identification and treatment of these cases. Nevertheless, no uniform opinion exists on handling these cases: Some researchers automatically recommend the exclusion without considering other factors; however, they view it as a kind of insurance policy, where the premium paid for the exclusion is comparably low to the distortion of results from data collection error. This assumes that a model or analysis fitting to the large majority of the cases is desirable. Thus, measures taken to achieve this goal are worthwhile. Others argue that decisions about the handling have to be taken on a per case basis and with respect to the phenomenon under study, the goal of the analysis and particularly the source of the extreme observations. Thereby, the information used to evaluate whether a case is erroneous or real is important and validation becomes an issue. Interestingly, the foundation of this dispute has apparently not been settled entirely since Bernoulli (1777). The scope of opinions also leaves room for the speculation that different research traditions (also called paradigms or schools) may influence this topic and the treatment of the issue.

Compared to marketing research, other areas of research have developed significant knowledge on the outlier phenomenon. Only little discussion could be found in marketing related literature even though dealing with data regularly. The next sections are therefore aimed to investigate the attitude of different research traditions towards this issue, the relevance for marketing publications, and finally to discuss handling alternatives and their consequences.

C. Practices in Marketing Research related to Extreme Cases

Marketing research strongly relies on empirical data analysis and survey research to test and gain scientific knowledge. As shown in the previous chapter, marketing literature has paid little attention to extreme and exceptional cases, whereas other disciplines exhibit a more detailed discourse. This section aims to shed light on (1) paradigmatic foundations and their consequences for the handling of exceptional cases, (2) the handling of these cases in premier marketing publications and (3) the positioning of the outlier issue related to similar phenomena from the sociology and psychology of science.

1. Paradigmatic foundations

Whenever researchers do their academic work, they make assumptions about the nature of the world (ontology) and how one can come to know it (epistemology). This worldview of a scientific community (Laudan 1977) is often referred to as "paradigm" or the "modus operandi of a scientific discipline" (Bortz and Döring 2002). Some philosophers of science argue that researchers who pursue different paradigms are, in a sense, living in different worlds (see e.g. "incommensurability of paradigms" by Kuhn 1970). Others explain that researchers are "born into orientations and paradigms rather than consciously selecting them" (Arndt 1985, p19). This ends with researchers suggesting that paradigms would not win over another by empirical and theoretical superiority, but because representatives of old paradigms die out (Bortz and Döring 2002). These paradigms are also referred to as research traditions which incorporate certain theories, metaphysical and conceptual assumptions. Their philosophical and methodological guidelines help to further develop the research tradition (Anderson 1983).

Paradigms also exist in the marketing discipline. One can see the beginning of the discussion with Bartels (1951) raising the question of whether marketing was a science or not; which was quickly answered in favour of the marketing discipline being a science. Others followed him in the discussion. Many "mechanisms" support the existence of paradigms and their diffusion. Particularly Ph.D. programs play an important role in the socialization process for researchers who are new to the field by emphasizing "model construction, hypothesis testing, data collection, and data analysis rather than a critical orientation" (Arndt 1985, p19). As the career progresses, some "guardians of faith" in gate-keeping functions such as journal editors, referees and editors help to preserve the status quo. The same happens more "naturally" at conferences, where people interact with persons they already know, for example, from their own country or even their own institution (Arndt 1985). Additionally, there are substantial rewards such as promotions and recognition if researchers stick to the research traditions.

Paradigms not only influence the way scientific knowledge is created, but also many other aspects, such as measurement theory or probability, which are crucial to the discussion of outliers. Savage and Ehrlich's (1991) book, for example, discusses measurement issues under different paradigms. As outliers arise in statistical calculations, the following section focuses on those orientations that allow for statistical analyses and may, therefore, not be as extensive as a textbook on research methodology (see e.g. Hunt 2003 for a marketing related or Smith 1998 for a more general account).

1.1. Idealism, (logical) empiricism and positivism

The late 1970s and 80s witnessed a philosophical debate in the marketing discipline on the "ideal" ontological and epistemological assumptions. Some key players advocated their schools of thought across a variety of journals. The transformation process of the marketing discipline in the 50s and 60s from a purely descriptive to a "rigorous and, therefore, acceptable" domain induced this discussion and gave rise to the positivist and logical empiricist movement (Easton 2002, p103).

At the time that these orientations received attention in the marketing discipline they were already preceded by a much longer history in the philosophy of science. Idealism, as one of the oldest philosophies of science, can be traced back to Plato and Antiphon. As a doctrine, idealism partly or completely refuses that one reality would exist as it is experienced without thought or ideas about it. In epistemological terms, idealism assumes that people only know their own ideas (images or representations from reality), but no objects outside of it. Consequently, all information and data are acquired through perception, but does not exist apart from it. Idealism is frequently juxtaposed to realism, which assumes absolute existence independent of one's knowledge about an object.

As a reaction to the "armchair thought-based systems of idealism", empiricism and positivism emerged (Easton 2002). Empiricism can be seen as a broad movement which mainly promoted experience and information through senses as sources of knowledge. Positivism is often considered as a part of this empiricist movement as it relies on facts accumulated through sense experiences and opposes metaphysical speculations. Positivism emphasizes the (positive) affirmation of propositions resulting in e.g. scientific laws.

Classical empiricism is rooted in the writings of John Locke (Russel 1961, p589) who departs from the human brain as a "white paper" which is then filled with experiences. According to Locke, no knowledge exists without reference to experiences; ideas arise from sensation (i.e. impressions from the external world) or reflection (ideas from the own mental operations). Knowledge of things is thus a perception of ideas that are in accordance or discordance with each other. Locke's "Essay Concerning Human Understanding" represents a milestone for the development of empiricism and historically an important starting point for the process initiating the Enlightenment. Locke's view opened the possibility that objects would not exist if they were not perceived – a question which was answered by bishop George Berkeley who argued that the omnipresent God would take over the role if they are not perceived by humans. John Stuart Mill's phenomenalism even rejected the existence of physical objects altogether and claimed their existence as perceptual phenomena and sensory stimuli only, created through inductive inference. During the Scottish Enlightenment, David Hume responded in his sceptical approach of empiricism to Berkeley and to the "continental" rationalist movement around René Descartes, who promoted innate ideas as source of human knowledge. He again emphasized sense experience as the source of all knowledge, while no beliefs about the world may be finally established by reason. According to him, all human beliefs exist due to an accumulation of habits and repeated experiences. Hume concluded that such things as belief in an external world and belief in the existence of the self were not rationally justifiable. Those beliefs were, nevertheless, acceptable as rooted in instinct and custom. Hume challenged the tendency of empiricists themselves to fall back upon ordinary and moral beliefs. He considered them to be the products of psychological processes which created problems in the ways that causal relationships should be identified (Smith 1998).

Most of the classical empiricists regard sense experiences to be the only source of knowledge: "All that we know is ultimately dependent on sense data" (Angeles 1992, p85). Reason - the rationalist's main source of ideas - does not provide any knowledge to an empiricist and can thus not lead to superior knowledge either. Reason may inform about the relation between ideas, but does not allow inference about the external world. Hume, for example, groups "all the objects of human reason or inquiry" in either "relations of ideas" such as Geometry, Algebra, Arithmetic and Geometry or "matters of fact". In his eyes, intuition and deduction can offer knowledge of necessary truths in the former areas, but such knowledge is considered as relations of ideas and not as substantive knowledge of the external world (Hume 1955, Section 4, part 1, p40). Substantive knowledge can only be gained through sense experience. This separation into "trivial" (a priori, prior to experience) propositions and substantive ("real world experienced" or "a posteriori") knowledge can already be found with Locke and yielded in the analytic/synthetic distinction in logical positivism, where "analytic truths were true by definition, without any need to refer to experience, whereas synthetic truths could only be known through observation" (Smith 1998, p60).

The strong focus on observability and non-relevance of reason also reinforced empiricists' doubt about everything beyond the physical world. Starting with Locke's "Essay Concerning Human Understanding" (1690/1990), empiricists attack the existence of metaphysical arguments and thus deny the existence of objects that cannot be established through direct observation (Hunt 2003, p29; Smith 1998, p59).

Hume was one of the first philosophers to recognize the problem of induction inherent to many empiricist philosophies: "A thousand observations of some event A coinciding with some event B does not allow one to logically infer that all As coincide with Bs." He concluded that there is no certainty that the future will resemble the past, for example, if we experienced two events in a simultaneous way several times in the past and would thus conclude that this would happen again. Hume suggests that "correspondence of constant conjunction of two observable events, sustained over time, is a contingent occurrence and does not imply any necessary relationship between the two variables, nor is it based upon their properties." (Anderson 1983, Smith 1998, p64). The problem of induction was later taken up by Karl Popper and represents a substantial part of his works.

Overall, classical empiricism mainly claims that knowledge may only arise from sense perceptions per se or mixtures and combinations of them and largely rejects anything beyond the observable. This includes particularly theological (except for Berkely) or metaphysical aspects. Hume (1777, p165), for example, rejected god or the self because of the lack of the impression from which it is derived. Reading the classical empiricist account, there does not seem any way for an idea to get into an individual's brain except by being related to experience. Thus, ideas which are not related to experience are considered not verifiable and consequently empty. Such verificationism is made explicit in logical empiricism and regards statements which cannot be verified as meaningless. In contrast, classical empiricism includes a somewhat weaker notion of verificationism and regards such statements as "empty" or simply inexistent. Such non-verifiable statements may include the relationships between the ideas that are discussed by many empiricist philosophers.

Two problems emerge with classical empiricism: First, perception is regarded as largely subjective, which means that individuals acquire their own, private experience, which renders the existence of objective (i.e. intersubjective) knowledge about – for example – commonly perceived objects or people impossible (e.g. Hume). This problem was solved by many philosophers by adopting subjective idealist or scepticist ideas. Second, the problem of induction is inherent to classical empiricist thought. In terms of content, classical empiricists do not discuss general sentences, but rather focus on pragmatic daily experiences. In all, the positions of the individual empiricist philosophers differ substantially, which renders a classification in empiricism categories difficult (see e.g. Kenny 1986 or Loeb 1981 for a discussion).

19th century positivism is frequently considered to be the first of three positivist époques, followed by the logical positivist movement – also known as logical empiricist of the Vienna Circle (e.g. Ayer, Carnap) and the standard positivist account (e.g. Hempel) (Outhwaite 1987, p5-8). Empiricism, in general, is often used synonymously for positivism and shares many assumptions. Some of these assumption are interpreted less strongly in postpositivist time. One can consider positivism as a "specialized" part of a major empiricist movement (Smith 1998).

In a nutshell: Positivism claims that authentic scientific knowledge must be based on sources which lay in the positive affirmation of theories through strict scientific method.

This first generation of positivists was much concerned with progress. Auguste Comte (1842), for example, aimed to reconstruct human knowledge by replacing the “divine truths”, which had collapsed through the enlightenment process, by using a “natural science” view on society – labelled “social physics”. He believed that scientific laws could be identified in much the same way as physical laws of motion and light. Using the laws of social statics and dynamics, Comte illustrated conquest and territorial annexation as well as the promotion of the white race as representatives for humanity using a “scientific” and thus ”objective” and value free language, while his cultural values in reality influenced his works. The contribution of Comte’s positivism can be seen in the assertion that *facts can be studied independently of values* (Smith 1998, p82). Herbert Spencer based his ideas on Darwin’s theory of evolution and Jeremy Bentham’s way of measuring human progress through the calculations of utility (i.e. satisfaction of human preferences) (Hunt 1991a, p251). According to Spencer, inequalities in society can be regarded as the consequences of competitive processes on the market, where those at the top are the fittest. As with Comte, Spencer’s work can be seen as the elimination of value statements by rationalization the eugenicist concerns with racial degeneration rather than with the objective study of facts. Thus, cultural values of time and place flow into academic writing while striving for “objective knowledge”. Positivist elements can also be found in Émile Durkheim’s philosophy. He promoted the idea that each discipline uses its own distinctive objects of analysis, which determine the way a discipline is constructed. These objects could be studied in the same way as the natural sciences, as “social facts”. He, for example, established empirical regularities (correlations) of suicide rates for different countries and regions across different social groups, as well as over time. He linked them to, considerable problems in trying to “establish relations of cause and effect between things which tend to correspond in a regular way.”(Smith 1998, p84).

As can be seen in the above-mentioned examples, positivism emerged as a response to the disorder and chaos within European societies at the time, and used the production of objective knowledge as a basis for reconstructing social life. *Thereby, positivists relied on empirical data derived from sense experience. Social facts and scientific laws became the common feature of social scientific practice, while influences from values were ignored.* These law-like relationships (Easton 1998, p81) are comparable to laws in natural sciences (e.g. physics) and aimed at progressing science (from divine knowledge). Established “facts” receive “truth” status. As with classical empiricism, facts arise through a verification process. Thereby, Comte did not explicitly discuss the meaninglessness of unverifiable sentences, but rather the lack of interest in them as they cannot be verified which renders such sentences useless. *To identify law-like relationships between constructs, constant conjunction or regular patterns of events allow to unveal these universal propositions.*

Related to the field of marketing: In the 1950s and 1960s, the discipline was transformed from a "descriptive and qualitative" to a rigorous and quantitative discipline, paralleled by "justifications that drew heavily on the positivist approach to research and theory development" (Easton 2002).

<u>Logical positivism and logical empiricism</u> emerged from the "Vienna Circle" in the 1920s to the 1950s and is also frequently called logical positivism, even though these two views are not completely identical. It is amongst the most influential scientific positions of the 20th century and aimed to create an objective account of human knowledge by using formal logic. Logical positivists combined a rationalist and an empiricist view by demanding a distinct separation between formal and empirical components in science. The previous metaphysical opposition of rationalism and empiricism was transformed into a linguistic distinction between analytic and synthetic truths (Hjoerland 2004, Hunt 1991a p268).

As their central element, logical positivists adopted the logic proposed by Bertrand Russel and Alfred Whitehead (Principia Mathematica) and Wittgenstein's verification theory of meaning (Brown 1977, Howard and Sheth 1969), which postulates that statements or propositions are cognitively meaningful only if there is a finite procedure to determine whether they are true or not (i.e. verificationism). A.J. Ayer extends verificationism by assigning "every cognitively meaningful sentence to one of two categories: Either it is a tautology, and as such true solely by virtue of the meaning or its terms and provides no substantive information about the world; or it is open to empirical verification." (Markie 2004) If empirical statements have been verified, they are considered 'universally true' which means that they are universally applicable. "There can be no a priori knowledge of reality. For… the truths of pure reason, the propositions which we know to be valid independently of all experience, are so only in virtue of their lack of factual content…. [By contrast] empirical proposition are one and all hypotheses which may be confirmed or discredited in actual sense experience." (Ayer 1952, p86 and p93-94 in: Markie 2004).

Such universality of truth statements were central to positivist thought, though, logical positivists dealt more carefully with it compared to 19^{th} century philosophers. As with other empiricists, logical positivists also used an inductive approach, which means collecting observations first and then building theories thereupon to explain them, leading to the problem of induction (Easton 2002). "…no finite number of empirical tests can ever guarantee the truth of a universal statement" (Anderson 1983, p19). To address the problem of induction, Carnap (1936; 1937) substituted the universality of truth statement with "gradually increasing confirmation" (Carnap 1953), a sort of asymptotic view of truth. "If verification is taken to mean the complete and definitive establishment of truth," then universal statements can never be verified. However, they may be confirmed by the accumulation of successful empirical tests (Carnap 1953 in Anderson 1983, p19f). In the marketing literature, Hunt (1976) suggested modifying the logical positivist view with the narrower vision of logical empiricism. *Several researchers observed that logical empiricism was the predominant paradigm for theory specification, theoretical progress and methods used in academic marketing research in the 70s and 80s, even though the philosophy of science had abandoned it over the previous decades* (Anderson 1983, Arndt 1985, Desphande 1983). Arndt (1985) relates this fact to "the applied nature of the discipline and to the 'channel captain' perspective, viewing marketing phenomena through the eyes of the corporate executive".

In terms of development of scientific knowledge, logical empiricism relies heavily on the analysis of empirical data. Peter and Olsen (1983, p118), for example, insist that "… only those knowledge claims that are based directly on experience (i.e. empirical observation) are considered important, useful and/or scientifically meaningful". These claims are consistent with the logical empiricists led by Carnap (1956), Hempel (1965) and Nagel (1961). Carnap (1946) for example describes the process as follows: "In order to test the theory we derive predictions with its help; then we observe whether and to what extent the facts bear out the predictions and these results are taken as the basis for the judgement of the theory" (p520). *Logical empiricism starts with the observation of reality and if the data are consistent with the hypotheses, a confirming instance is identified. Strict logical empiricism does not allow unobservable concepts and variables to be part of scientific explanation (Easton 2002). Scientific progress arises through the "accumulation of multiple confirming instances obtained under a wide variety of circumstances and conditions" (Anderson 1983, p18f*). In comparison to classical empiricism (e.g. Bacon) – where generalizations were induced from observations – logical empiricism separated the context of discovery from the context of justification (see Reichenbach 1938) and focused exclusively on the justification of hypotheses, laws and theories (Hunt 1991a). Overall, one can summarize that "logical positivism combines the method of induction with the verificationist approach to testing scientific explanations, whereby scientists seek out similar observable conditions to demonstrate the validity of the general law established so far" (Smith 1998, p100).

Truth was distinguished into synthetic and analytic truths where "analytic truths were true by definition, without any need to refer to experience, whereas synthetic truths could only be known through observation" (Smith 1998, p60).

One major critique related to the logical empiricist view is rooted in the reliance on observation that at first seems objective, but is later more and more considered to be theory laden. Additionally, logical empricists assume a (artificially) closed system perspective through which scientific explanation is possible. The empirical approach proposed by logical empiricism can be described as such:

> *"Where the sampling of phenomena was inevitable, statistical inference could be used to provide gradually increasing confirmation in a probabilistic sense" (Easton 2002, p104).*

In that way, logical empiricists approached Humean scepticism (i.e. that people are able to make causal determinations only in a subjective fashion that does not exist in the real world, but only shaped by repetition in the human brain) as both a problem to be solved and a problem that could be solved through the development of the logic of probability. Consequently, positivist and logical empiricist "laws" are reflected in regularity and co-occurrence as well as co-variation, but no causations (Hunt 1991a, p289). These relationships are faced with problems of sampling and measurement. Therefore, a 100% correlation cannot be expected.

In this way, the paradigm provides some rationale for errors and non-confirmation of the laws under study. Cases that do not confirm the pattern are considered to be errors in statistical methods irrespective of their validity in terms of content.

One can extrapolate that disturbing outliers are also easily grouped in this category and conclude that errors are not particularly desirable and interesting. Thus, they may be eliminated or transformed and do not require reporting as the focus lies on the discovery of relationships. While logical empiricism renders this argument very straightforward, other empiricist philosophies are based on similar assumptions (e.g. verificationism). Thus, outlier treatment can be seen in a similar light. "Rules have been evolved which help researchers decide whether the correlations, in whatever form calculated, are significant" (Easton 1998, p81). If they have not been rejected, yet several replications have been successfully performed, these rules receive gradual confirmation. Explanations of complex patterns within the data consist of summaries of relationships among "artificially" created variables (Easton 1998).

Interestingly, many statistical procedures such as Fisher's Analysis of Variance or Pearson's correlation and regression analysis have emerged from the work of the early empiricists. Both streams build on inductive inference (i.e. inferring from a sample to a whole population). According to Fisher, repeated refutation of a null hypothesis or indirect proof of effects fosters knowledge creation. As a criterion for rejection, Fisher (1925b, 1925a) proposes the very popular falsification criterion ("$\alpha < .05$"). Even though a critical rationalist, Popper (1989) also urges for such a criterion, but leaves the decision for a demarcation criterion up to the researcher. Otherwise, probability statements would be purely metaphysical because of a lack of falsifiability (see Bortz and Döring 2002).

The standard positivists such as Carl Hempel modify the approach taken by the logical positivists and stress the logical deduction of conclusions from generalizations or universal laws. The starting point, thereby, is theoretical, even though this theory may have been established through induction. While the logical positivists argue that statements are only meaningful if they describe the reality "as is", Hempel recognized that "the empirical content of a theoretical statement could only be considered within its interpretive system" (Smith 1998, p103). Truth is reached when a scientific theory predicts a suggested outcome accurately. Examples of this perspective include Skinner's behaviourism, which focuses on the study of changes in behaviour by humans or animals as a reaction to stimuli.

Post-positivists recognize the critique on the positivist approach and partially acknowledge the critique including the independence of researcher and researched person (Robson 2002, p27). Post-positivists accept that theories, hypotheses, background knowledge and values can influence the researcher (Reichard and Rallis 1994) even though they commit to objectivity and try to correct for those biases, for example. Post-positivists also postulate the existence of one reality, but agree that it can be known only imperfectly and probabilistically. "Post-positivist researchers can be viewed as recognizing, sometimes reluctantly, that the battle for positivism has been lost, but are still hankering after the mantle of respectability and authority that it conferred…. It is argued that they can find their salvation in critical realist approaches" (Robson 2002, p27).

Assumptions of empiricist and positivist approaches

In the previous sections, *verificationism could be identified as one major commonality across different forms of empiricist and positivist paradigms*. Verificationism, nevertheless, has often been used in a confirmationist and justificationist way in order to provide evidence with which social scientists feel comfortable (Smith 1998). This may also include the introduction of conditions (e.g. the exclusion of disturbing cases), if they are not covered by probability calculus.

Additionally, empiricist paradigms assume that the universe consists of atomistic, discrete and observable events and the social sciences are no exception to that. Thus, social objects may be studied in the same way as nature (Smith 1998, p77) and no difference is made between essence and the phenomenon. Metaphysical statements that reach beyond the observable are largely rejected, although unobservable statements and concepts are accepted by some logical empiricists (Easton 2002). Particularly in positivism, progress is considered an important aspect of science. Positivism and logical empiricism contribute the strong distinction between facts and values and the nature of scientific laws to the assumptions.

Criticism against different streams of empiricism focuses on the impossibility of a methodical structure – even in "exact sciences" such as physics – without any a priori synthetic sentences. Furthermore, all efforts to conceive a basis for science on uninterpreted data were shown to be impracticable. Thus, theory's dependence on experience is widely accepted today (Mittelstrass 2004).

Table 16 details the key assumptions described and used in this section.

	Definition
Naturalism	Positivists are committed to naturalism, the idea that it is possible to transfer the assumptions and methods of natural sciences to the study of social objects, often referred to as the 'unification of method'.
Phenomenalism	Phenomenalism is the assumption that only knowledge gained through observed experience can be taken seriously. If something cannot be directly experienced, it is said to be metaphysical – beyond our physical senses.
Nominalism	Nominalism shares with phenomenalism the argument that concepts must be based upon experience, but it also asserts that concepts have no use other than as names. Words are seen as pure reflections of things. It is, of course, very difficult to do this because the words we use are usually far more than simple descriptions.
Atomism	Atomism is a particular approach to the definition of objects. Atomism states that the objects of scientific study are discrete. That is that the objects cannot be broken down into any smaller parts. These objects act as the foundations of a scientific study. Collective objects are thus the sum total of their smaller atomic components.
Scientific Laws	The purpose of science is to develop laws. To develop a scientific law, you start from the observation of a particular set of objects and look for regularities. The regular appearance of two or more things together or in some kind of sequence can be called an empirical regularity. This is sometimes described as a constant conjunction of events; one then explores whether the same regularities occur in other similar circumstances. A scientific law is a general statement that describes and explains empirical regularities that occur in different places and at different times.
Facts/ Values	Facts and values are seen as distinct. Only facts can be regarded as scientific. Facts can be empirically verified, measured and explained by reference to observational evidence. Values involve subjective assessments and claims about what ought to be. Thus, values cannot be observed, measured or explained.
Source: adapted from Smith 1998, p76	

Table 16: Assumptions of positivist and empiricist philosophies

1.2. (Critical) Rationalism, relativism and realism

Logical positivism and empiricism were shown to be illogical by David Hume, who especially argued against the principle of induction: "A thousand observations of some event

A coinciding with some event B does not allow one to logically infer that all A's coincide with B's". Rationalism, as a realist philosophy, is often considered the opposite to empiricism. Rationalists depart from humans having "innate knowledge", meaning that from birth, humans start with all knowledge available to them. Innate knowledge is considered part of human nature. Philosophers, though, differ in their opinion on the sources of this knowledge: Some propose an earlier existence, others God and others attribute it to nature through the process of natural selection. They, nevertheless, reject sense experience, intuition and deduction for the creation of innate knowledge.

Classical rationalists such as Descartes, for example, suggest that "all knowledge is certain and evident cognition" and when we "review all the actions of the intellect by means of which we are able to arrive at a knowledge of things with no fear of being mistaken, [we] recognize only two: intuition and deduction" (Descartes in: Markie 2004). Classical rationalism assumed an existing order in the universe, which can be mapped by human knowledge regardless of a particular point of view. Similar to empiricism, rationalism shares the view that methods are independent of the concepts, theories and points of view of the researcher (Hjoerland 2004). One gains propositions in a particular area by intuition or deduction from intuited propositions. Deduction is thereby a process, where one derives conclusions from intuited premises through valid arguments. This is also often referred to as "a priori". Concepts are thereby part of the rational nature, sense experiences are regarded as triggering those innate ideas to consciousness, but experience provides neither content nor information. Whether intuition leads to the acquisition of truth or not is answered in a similar way by rationalists. Leibnitz and De Spinoza are other well-known representatives of this stream of thought (Hunt 2003, p31).

While classical rationalists argued that science should be built on rational intuition as a sole method of research and, therefore, rejected empirical data per se (Hjoerland 2004), empirical hypothesis testing is considered a key element of critical rationalism, most prominently advanced by Karl Popper (1962, 1966). Critical rationalists assume the existence of a real world that is independent of human knowledge about it. Human beings are limited to their way of perception Hence, there is no final certainty as to whether these perceptions correspond with reality. Following the arguments of the critical rationalists, human beings can gain knowledge positively by experience either by turning observations away from existing conjectural theories pertinent to the observations or from underlying cognitive schemas which unconsciously handle perceptions and use them to generate new theories. Critical rationalists thereby are not interested in the existence of those theories in reality, but are content with their empirical observability. As these theories are advanced in comparison to what was observed, they are not logically induced and may therefore be wrong. Many people induce theories from observations because this way is frequently marked with success. One frequently used example is: If induction were possible, the sun would rise each morning because I get up (The sun was always rising when I got up). Popper dismissed the programme of the logical empiricists that attempted to find a probability formulation for the "degree of confirmation" of a law-like generalization. Popper also disagreed with the verifiability principle as a method of demarcation, as law-like generalizations cannot be verified in the sense of being deductively true (Hunt 1991a, p291).

Popper proposes "conjectures and refutations" as an approach to overcome the problem of induction. In contrary to positivist and empiricist thought, which does not assume any a priori knowledge, Popper (1972, p344) in critical rationalist manner acknowledges that "observation always presupposes the existence of some system of expectations". *In Popper's eyes, science starts progressing when observations collide with existing theories and preconceptions, called a scientific problem (Anderson 1983). To solve this problem, a theory is suggested and the logical conclusions of the theory (hypotheses) are rigorously, empirically tested. In Popper's logic, the origin of a statement or a hypothesis is unimportant, as long as it can be submitted to an empirical test.* Compared to logical positivists, who saw statements not describing the real world as meaningless, Popper regards statements that are useless for theory generation as "non scientific". *"The falsifiability of a system is to be taken as a criterion of demarcation"* (Popper 1959, p40) *whether a statement is scientific or not.* Pseudo-scientific statements already include the conclusion in its premises and thus, no falsification is possible. Science, for Popper, is no more the quest for universal and objective truth, but the careful and systematic application of the scientific method – meaning the construction of statements about the world and testing them. Therefore, "scientists should find ways of disproving their working hypotheses and theories" (Smith 1998, p108). According to Popper, the focus on refutation rather than confirmation prompts scientists to think harder about what they do. Hypotheses which survive attempts of refutation are seen as "corroborated", but not confirmed. Scientific progress is reached through testing and attempted falsification of hypotheses and – if necessary – by modified or new ones (Smith 1998, p109).

Following Popper, theories cannot be confirmed, but only corroborated. If they survive severe falsification, they are said to be corroborated and are tentatively accepted, but the degree of corroboration cannot be associated with a probability (Hunt 1991a, p291). Probability is acceptable to Popper (1962) as an aid to prediction, but not for the assessment of the value of a particular theory (Cowles 2001, p30). In case a theory's predictions are falsified, the theory needs to be rejected (Anderson 1983). Popper emphasizes the fact that a universal hypothesis can be falsified by a single negative instance. Conjectures and refutations, therefore, represent the basis of scientific progress in the view of Popper's falsificationism (Popper 1962) and can be seen as a continual process. *If deductively derived hypotheses are shown to be false, the theory is taken to be false and the problem of induction as in positivist and empiricist thought should be avoided.* Exhibit 6 juxtaposes logical empiricist and critical rationalist hypothesis testing.

Many researchers have been attracted to Popper's research programme, especially through "falsification's image of science as a rational and objective means of attaining truth" (Calder, Phillips and Tybout 1981). It is "not the possession of knowledge of irrefutable truth that makes the man of science, but his persistent and recklessly critical quest for truth" (Hunt 1991a, p292). Critics of Popper's view include Duhem (1953) who argued that it would be impossible to finally oppose a theory. The test situations depend on much more than just the theory that is under investigation (e.g. measurement equipment), consequently, an empirical test involves assumptions about initial conditions, measuring instruments and additional hypotheses. The Quine-Duhem (Quine 1953, Duhem 1962) thesis points at the problem related to background hypotheses. If testing happens in isolation, background assumptions might be counteracting. Thus, any outcome can be rationally questioned, mistrusted and explained away by ad hoc hypotheses that change the background conditions. Falsification is therefore called dubious and ambiguous (Cook and Campbell 1979). The Quine-Duhem thesis is also frequently referred to in relativist philosophies.

The principle of falsification represents a central element in hypothesis testing under a critical rationalist approach. Popper considered the query for falsification as crucial to the progress of knowledge, whereas the process of developing new ideas and scientific problems itself is characterised by trial and error as well as creativity and chance. He urges the researcher to seek for a refutation, or a counter example in order to constitute a serious empirical test (Popper 1966 p54, Johansson 1975 p70). *Under this perspective, extreme and disconfirming cases offer the opportunity for disconfirmation and may therefore progress theoretical development.* One can assume that a researcher influenced by this paradigm would be highly interested in discovering these cases and consequently report and discuss them in the context of the study before maybe arguing for a certain method of managing them.

<table>
<tr><td colspan="2"><u>Hypothesis testing under different paradigms</u>
Hypothesis testing takes a central role in positivist / logical empiricist as well as critical rationalist methodology. While discordance exists in marketing literature whether positivism or any other philosophy of science dominates (see e.g. Hunt 1991b against Easton 2002 or Hunt 1994), hypothesis testing is maybe the most admired research strategy (Greenwald, Leippe, Pratkanis and Baumgardner 1986) and very popular among researchers. At first glance, theory testing in both positivist and critical rationalist paradigm look quite similar: Hypotheses are developed and their predictions are then tested against relevant data. Nevertheless, some important differences exist.</td></tr>
<tr><td><u>Logical empiricism / positivism (e.g. Carnap 1953)</u></td><td><u>Critical realism (e.g. Popper 1972)</u></td></tr>
<tr><td>No knowledge exists before testing.</td><td>Observation always presupposes the existence of some system of expectation.</td></tr>
<tr><td>Theory gradually gets confirmed through successful replications.</td><td>A theory can only be falsified. The survival of severe falsification attempts make theories temporarily accepted.</td></tr>
<tr><td>Negative replications do not necessarily impact the validity of the theory.</td><td>A single negative replication makes the theory invalid.</td></tr>
<tr><td>Truth exists and manifests itself in knowledge.</td><td>Science means the persistent and recklessly critical quest for truth.</td></tr>
</table>

Exhibit 6: Comparison of hypothesis testing under different paradigms

Relativism

Relativism emerged as a counter movement to the standard view of science. Historically, sophist philosophers in Ancient Greece (e.g. Protagoras) can be seen as the ancestors of relativist thought. More recent representatives are Paul Feyerabend and Thomas Kuhn (Hunt 1991a, p321). Kuhn, for example, argues that truth of propositions or their existence of an entity is relative to the paradigm employed. Feyerabend - overall - promotes anarchy in the philosophy of science ("anything goes"), leading scientists to ignore philosophical guidelines as they limit progress. In other stances, he details strong relativist ideas by regarding science as "just one cultural tradition among many" along with religion or astrology (Feyerabend 1978 in Robson 2002, p21).

Relativism can be considered as a group of philosophical positions using principles that do not possess universal validity for the foundation of sentences. In an epistemological form, this stream suggests that *knowledge is contingent to internal factors within a scientific discipline – such as theory ladenness of observation or methodological criteria for the evaluation of empirical theories (cognitive relativism). On the other hand, this stream suggests the influence of the social situation on the subject and thus the assumptions of the object of investigation (sociological relativism, also known as scientific constructivism)*. The former relies on the validity of other theories and standards of evaluation under different paradigms for the confirmation of theories, the latter on the acceptance through social, economic and institutional norms (Krausz and Meiland 1982, Muncy and Fisk 1987, Mittelstrass 2004).

In its most extreme form, relativism postulates that there is "no external reality independent of human consciousness"; there are only different sets of meanings and classifications which people attach to the world (Robson 2002, p22). This means that truth is relative to something other than itself – such as subjects, time or situations – and no objective truth can be identified (Hunt 1990). Some positions even go further and emphasize that there is no reality, often referred to as anti-realism. Critics of universal relativism often argue that relativist positions must then be held against each and every proposition about truth, which implies that they do not have more truth status than their opposite. This would lead every relativist into permanent self contradiction, as the act of communication would stand in contradiction to the content of the communication. Thus, relativist positions often differ in their extremeness and relativism can also be found as an influence in other philosophies of science (see e.g.Hunt 1990 or Anderson 1986 for a related discussions).

Reality can only be constructed by means of a conceptual system, and hence there can be no objective reality because different cultures and societies have different conceptual systems

(Robson 2002, p22). This form has been challenged by many researchers. One of these attacks was launched by Davidson and Layder (1994, p26):

> *"It is possible to accept that neither natural nor social scientists ever rely purely on observation and that their observations are never completely detached from their pre-conceived beliefs and theories about the world without having to argue that empirical observations are therefore useless, or that they are invariably disregarded by scientists, or the science is merely the subjective process by which they got about confirming their own prejudices. It is essential to recognise that empirical observations are not the be all and end all of scientific research, and that a commitment to natural science method and procedures does not offer the final, ultimate, reliable and objective way to obtain knowledge, without abandoning all belief in reality or all hope of advancing human knowledge."*

Relativism per se may not appear frequently in marketing literature; it can often be found under "constructivist, naturalistic or interpretive" approaches and frequently linked to qualitative, "non-statistical" analyses (see e.g. Goulding 1999 for a summary). Constructivism, in the form of a theory of knowledge, suggests that all knowledge is constructed and thus contingent on convention, human perception and social experience (e.g. Mir and Watson 2001). Critical relativism – as one version of the relativist philosophy of science in marketing science – postulates that no single scientific method exists, but all depend on those who practice science, their beliefs, values, standards, methods and cognitive aims. Knowledge production is consequently impacted by the broader cultural milieu to which it relates (Anderson 1986). The value of such claims needs to be evaluated in the light of their production and methods of justification. Empirical testability alone is therefore too limited in terms of a scientific perspective as also pursued by non-scientific disciplines (Anderson 1986). Whether a new theory is accepted or not depends on the extremeness of relativism. Moderate relativism attributes to social factors a co-decisive role in the evaluation of theories, filling up the freedom left by other scientific test criteria. Extreme relativism relies entirely on social factors (Schofer 2000).

Critical relativism rejects the existence of a single reality to be discovered “out there” by the scientific method (Olson 1981). The critical relativist will resist the assertion that science is capable of revealing or even approximating this "reality" (Laudan 1981). Critical Relativism therefore advocates competing research programmes in order to explore a phenomenon, each with its own advantages and problems (Anderson 1986). This stream is sometimes claimed to be self-refuting, as no “truth” can be claimed for some positions, because they are relative at the same time. Though, counter arguments exist (see e.g. Hunt 1990). Anderson (1986) makes the point that the existence of relativism makes researchers aware that all results are contingent on a certain époque in time, research school etc. Qualitative methods dominate these research tradition, even though Feyerabend promotes its anarchical “everything goes” approach.

Outliers also exist in qualitative data, as shown in section B.2.3. Besides the “technical” details presented there, a relativist or constructivist researcher may be interested in the contingencies of such cases which serve also as a reference for interpretation. In a relativist view, data can only be interpreted in the circumstances they were created. Thus, the researcher may at first re-interpret the data with the subject or object used for data collection. Thereby, a deeper understanding of the “milieu” (i.e. beliefs, values, standards and methods) in which knowledge and the outlier have emerged, are important. This also helps to determine whether this case is outside of current knowledge of the discipline and thus not supported. This view can also be applied to Feyerabend’s suggestion that qualitative as well as quantitative research is possible (“everything goes”), as long as they can stand the test. *Whether an extreme or outlying case is then considered interesting under a relativist philosophy depends on the social, economic and institutional circumstances. In order to become fact, and thus interesting for others, the nature of the outlying case needs to be “negotiated”, meaning that it be discussed and accepted by the (scientific) environment.*

Table 17 summarizes the commonalities of most relativist approaches.

Features of relativist approaches
1) Scientific accounts and theories are not accorded a privileged position; they are equivalent to other accounts (including lay ones). Different approaches are alternative ways of looking at the world and should be simply described, rather than evaluated in terms of their predictive power, explanatory value or truth value.
2) It is not accepted that there are rational criteria for choosing among different theoretical frameworks or explanations; moral, aesthetic or instrumental values of conventions always play an essential part in such choices.
3) Reality is represented through the eyes of participants. The existence (or accessibility, which has the same consequences) of an external reality independent of our theoretical beliefs and concepts is denied.
4) The role of language is emphasized, both as an object of study and as the central instrument by which the world is represented and constructed.
5) The importance of viewing the meaning of experience and behaviour in context, and in its full complexity, is stressed.
6) The research process is viewed as generating working hypotheses rather than immutable empirical facts.
7) The attitude toward theorizing emphasizes the emergence of concepts from data rather than their imposition in terms of a priori theory.
8) Qualitative methodologies are used.
Source: Robson 2002, p25

Table 17: Features of relativist approaches in social sciences

Realism

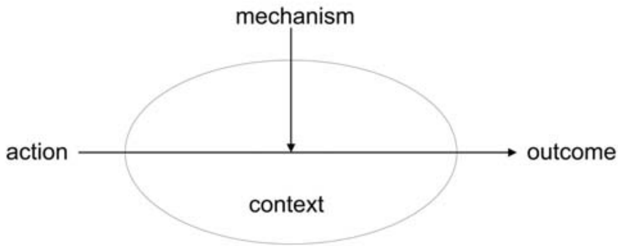
Scientific realism roots in the critical, dynamic, empiricist, fallibilist and evolutionary epistemologies of the 19th century, such as Peirce's pragmatism and Engel's dialectic materialism. In the 20th century, logical positivism was followed by philosophers developing scientific realist thought, such as Karl Popper or Rom Harré (Burkhardt and Smith 1991, p761). Scientific realism embraces various philosophical positions about scientific knowledge. They all hold that there "exists a reality independent of human minds and scientific theorizing is a good method for gaining knowledge about the mind independent reality" (Burkhardt and Smith 1991, p761).

As one of the earliest forms, naïve realism emphasizes the idea that things are essentially the way they are perceived. Objects have properties such as size, colour or shape, which can be perceived correctly. While many empiricists (e.g. Locke with his "prima quality") held naïve realist ideas, this stream was particularly criticized by philosophers who pointed at conflicting appearances (e.g. Bertrand Russel): The same object may appear different to various people or over time (see e.g. Hunt 1990). Thus, this stream is not further discussed here.

"The majority of philosophers of science profess to be scientific realists" (Causey 1979, p192) and, according to Hunt (1990), *most research programs in marketing are at least consistent with scientific realism.* Researchers seldom explicitly declare their position, but "much marketing research seems implicitly to assume a realist perspective" (Hunt 1990). Scientific realism accepts that reality exists independent of our knowledge of it. Reality is complex, but not random; relationships can be found which make it worth trying to study and interpret reality. Observation is important, but in contrast to positivism, observations are subject to reinterpretation (Robson 2002, p34). *The realist accepts that knowledge is a social and historical product and that our interpretations of "facts" are influenced by the theories we use for guidance* (Sayer 1992). Truth is possible from a realist point of view and the self-corrective and truth-producing method of science is the best way to approach it. Nevertheless, even the best theories and observations which refer to real objects and processes are fallible and correctible. *The use of systematic methods of science at least makes it highly probable that the scientific community will eventually reach truth-like or approximately true information about reality.* As a doctrine about scientific theories, scientific realism claims that theories are true or false attempts to describe reality. Talking about the existence of unobservable entities, they may receive indirect support from the empirical success of the theory (Burkhardt and Smith 1991, p762).

Realism accepts that there are fundamental differences between natural and social phenomena. Thus, different methods are needed for different subject matters. Nevertheless, overlap is not impossible (Robson 2002, p35). Realist explanations work in terms of mechanisms as shown in Figure 3. Robson (2002) uses the example of gunpowder for explanation:

> *"Does gunpowder blow up when a flame is applied? Yes, if the conditions are right. It doesn't ignite if it is damp, or if the mixture is wrong, or if no oxygen is present, or if heat is applied only for a short time. In realist terms, the outcome (explosion) of an action (applying the flame) follows from mechanisms (the chemical composition of the gunpowder), acting in particular contexts (the particular conditions which allow the reaction to take place)" (Robson 2002, p30).*



Source: Robson 2002, p31

Figure 3: Representation of realist explanation

Studies carried out under strict laboratory conditions (context) allow specifying when certain conditions (mechanisms) apply or not, even though this is sometimes questioned in some physicist research fields. Realists nevertheless recognize that social phenomena happen in real life, which is rarely a closed system and more than one mechanism might be operating at the same time. They agree that these mechanisms can be different from context to context. Human behaviour can only be understood within different layers of social reality: Therefore, the context in which these mechanisms are working is essential. This view is consistent with the social reality happening in an open system where a high degree of closure is seldom possible (Manicas and Secord 1983). Thus, probabilities and tendencies are in use. To complete the picture, not only the context where an effect has been successfully shown is of interest to the researcher, but also where this effect has not been present (Robson 2002). Even though the degree of closure needed to establish perfectly predictable regularities is rarely available, mechanisms still apply.

The strongest argument in favour of scientific realist thought lies in the "no miracle argument"; meaning that if a theory is capable of predicting a new empirical regularity in a correct manner, or if a theory allows the uniform description of previously distinct phenomena, then this theory is considered to describe the processes in a correct way (i.e. prediction is no miracle). This makes the theoretical entities exist in reality (Mittelstrass 2004, p507).

Realism has been employed in many forms such as "new realism", "critical realism", "scientific realism", "fallibilistic realism", "subtle realism" and "transcendent realism". Each version stresses particular features and the area is rather disputed. Consequently, no one of the protagonists, such as Roy Bhaskar or Rom Harré, is likely to agree with the summary displayed in Table 18 (Robson 2002, p29).

A realist view of science
1) There is no unquestionable foundation for science, no "facts" that are beyond dispute. Knowledge is a social and historical product. "Facts are theory laden".
2) The task of science is to invent theories to explain the real world, and to test these theories by rational criteria.
3) Explanation is concerned with how mechanisms produce events. The guiding metaphors are of structures and mechanisms in reality rather than phenomena and events.
4) A law is the characteristic pattern of activity or tendency of a mechanism. Laws are statements about things that are "really" happening, the ongoing ways of acting of independently existing things, which may not be expressed at the level of events.
5) The real world is not only very complex but also stratified into different layers. Social reality incorporates individual, group and institutional and societal levels.
6) The conception of causation is one in which entities act as a function of their basic structure.
7) Explanation is showing how some event has occurred in a particular case. Events are to be explained even when they cannot be predicted.
Source: Robson 2002, p32

Table 18: Features of realist scientific approaches

While scientific realism accounts for a "family of similar paradigmatic streams", "critical realism has attracted the attention of social scientists in recent years by offering attractive solutions to the problems associated with both, positivist and hermeneutic philosophies of science" (Pratschke 2003, p13f). At a broad level, a critical realist would "argue that the generation of theory implies critically evaluating the state of existing knowledge and building upon this to create new theoretical approaches" or "can be conceptualized as an umbrella term to refer to theory in a critical tradition including philosophy in the case of critical realism, feminism, racism, and postmodernism" (Burton 2005, p11).

In comparison to other forms of realism, a fallabilistic and critical realist postulates that "some of our perceptions may be illusions or even hallucinations … some of our perceptions may be true and others false or, alternatively, some of our perceptions are "more accurate" or "closer to the truth" than others" (Hunt 1990, p9). Table 19 enriches the description of this approach using Sayer's "signposts" for the critical realist thought, which is discussed afterwards.

Sayer's (1992) "signposts" characterizing critical realism
1) The world exists independently of our knowledge of it.
2) Our knowledge of the world is fallible and theory-laden. Concepts of truth and falsity fail to provide a coherent view of the relationship between knowledge and its object. Nevertheless, knowledge is not immune to empirical check and its effectiveness in informing and explaining successful material practice is not mere accident.
3) Knowledge develops neither wholly continuously, as the steady accumulation of facts within a stable conceptual framework, nor discontinuously, through simultaneous and universal changes in concepts.
4) There is necessity in the world; objects—whether natural or social — necessarily have particular powers or ways of acting and particular susceptibilities.
5) The world is differentiated and stratified, consisting not only of events, but objects, including structures, which have powers and liabilities capable of generating events. These structures may be present even where, as in the social world and much of the natural world, they do not generate regular patterns of events.
6) Social phenomena such as actions, texts, and institutions are concept-dependent. We therefore not only have to explain their production and material effects but to understand, read, or interpret what they mean. Although they have to be interpreted by starting from the researcher's own frames of meaning, by and large, they exist regardless of researchers' interpretation of them. A qualified version of (1), therefore, applies to the social world. In view of (4)–(6), the methods of social science and natural science have both differences and similarities.
7) Science or the production of any kind of knowledge is a social practice. For better or worse (not just worse), the conditions and social relations of the production of knowledge influence its content. Knowledge is also largely — though not exclusively — linguistic, and the nature of language and the way we communicate are not incidental to what is known and communicated. Awareness of these relationships is vital in evaluating knowledge.
8) Social science must be critical of its object. In order to be able to explain and understand social phenomena, we have to evaluate them critically.
Source: Sayer (1992, p5) in: Easton (2002, p104)

Table 19: Characteristics of critical realist thought

Signposts 1 and 2 thereby characterize how one can "come to know". *Thereby, critical realism accepts that perception and error in perception and data collection may exist and data can be interpreted differently as the world is in reality. "In particular, the realist claim is that accepting that successful scientific theories describe truly (or, nearly truly) the* *unobservable world best explain why these theories are empirically successful"* (Psillos 1999, p71). Spoken along the "no miracle argument", that which needs to be explained, the explanandum, is the overall empirical success of science. Successful theories are approximately true and must be accepted on these grounds. Nevertheless, it also supports the fallibilist doctrine, implying that no absolute certain knowledge is possible and that existing knowledge can be revised by further observation. Even theories that are false can still be valuable if they are closer to the truth than their competitors (Popper 1962). These observations may rise from the collection of new data, but also from in depth analysis of the existing data and, thus, the critical investigation of disproving cases.

Signposts 3 and 7 describe the development of knowledge as a somewhat continuous process and a social practice. *The context of the knowledge development influencing the knowledge produced* has already been outlined in section 1. From this section and signpost 8, one can derive the *suggestion to implement practices seeking disconfirmation for existing theories and practices.* One of these practices could be the study of disconfirming cases. Signposts 4 and 5 can be viewed particularly from the open systems perspective promoted by critical realist thought. When collecting data, particularly using highly structured instruments, one cannot rely on capturing all relevant mechanisms generating events. *Critically reviewing and potentially re-collecting data from extreme and deviant cases may help to build awareness of additional mechanisms and "background hypotheses" in place and, potentially, establishing conditions for, thus far, very general theories (e.g. Greenwald, Leippe et al. 1986).*

While publications by the early researchers in this area (e.g. Bhaskar 1975, 1998, Sayer 1992, Lawson 2001, Fleetwood 2001) have clearly rejected the use of statistical methods, other authors advocate their use (Olsen and Morgan 2005, Porpora 2001) in general and in particular structural equation modelling (Pratschke 2003), multiple regression analysis (Naess 2004) or logistic regression (Olsen and Morgan 2005) under this philosophy of science. The arguments against the use of statistical methods focus on the "black box" mode used by primarily logical positivist researchers (Olsen and Morgan 2005), while those in favour promote a very deliberate and epistemologically sound way of using such methods and interpreting the results. Most of the suggested statistical tools are related to the development and discovery of theoretical knowledge. Manicas and Secord (1983, p399f) propose the following formulation for the experimental use of statistics under the realist paradigm:

> *"In effect, the statistically significant difference between the mean performances of two or more sets of individuals so common to experiments is a kind of probability usage. What it amounts to is making a crude generalization by attributing individual variances to "error" when, in fact there are usually real differences between individuals. Another use of statistics is at the macro level, as in economics, demography or educational sociology.*
>
> *Do findings on such methods constitute explanations? This question has been frequently debated, but the answer depends on what one takes as scientific explanation. From the perspective of the new philosophy of science, however, they clearly do not constitute explanations. For that viewpoint, explanations must be based on a dynamic structure/process that has causal forces. Mere description is insufficient. Yet, this does not mean that statistical methodologies and probabilistic approaches are useless. Far from it. Often they are an earlier necessary stage in the process of description and discovery, and they may, for example, lead to the discovery of powers or competencies that are at first understood only at the macro level, but that later may be explained in terms of generative mechanisms."*

Thus, under this philosophy of science, the identification of extreme and deviant cases may be the point of departure for a new research venture. Bhaskar (1986 in: Pratschke 2003) argues that "a theory is preferable to another if it can either (a) identify and/or describe and/or explain a deeper level of reality; and/or (b) achieve a new order of epistemic (explanatory and/or taxonomic) integration, or at least show grounded promise of being able to do so" (p82). *Speaking in empirical terms, one might therefore assume that a theory including extreme cases would be considered superior to one that does not account for them*. Different to Popper's Falsificationism, "scientific theories are falsifiable in the sense that any given "counter example" will overturn them. Obviously enough, we do make some sort of forecasts about what will happen in open systems on the basis of scientific theories, but nothing that happens in an open system will of itself falsify a theory" (Collier 1994, p58). Therefore, deviant cases do no have the same power in terms of refutation as under a critical rationalist philosophy.

Table 20 summarizes the role of exceptional cases under different paradigms.

Paradigm	Relevance of exceptional cases
Positivism/logical empiricism	Empiricists seek to verify their empirical propositions. Disconfirming cases may lead to giving up propositions. Probability (logical empiricists) may help to "gradually confirm" propositions and take care of disconfirming cases. In practice, verificationism has been frequently interpreted as confirmationism (Smith 1998), which may explain the exclusion of such cases from analysis.
Critical rationalism	A single negative instance can falsify a theory – deviant cases can therefore have high importance and advance science. Their reporting and discussion is highly relevant.
Relativism/ Constructivism	Extreme and outlying cases need to be interpreted in the context they arise. Whether they are relevant or not is "negotiated" by the (scientific) community.
Critical realism	A theory including extreme cases would be considered superior to one that does not account for them. A negative instance (i.e. where the theory does not apply) is considered interesting for theoretical progress and does not provide reason to reject the whole theory. Particularly the conditions under which this negative instance arises are worthwhile studying.

Table 20: Treatment of exceptional cases under different paradigms

1.3. Handling outliers in the light of different paradigms

Different options to deal with outliers have been identified in the statistical literature in section B.2.3. The following paragraphs discuss these handling options proposed from a "technical" point of view in the light of different paradigms. This aims to shed light on the potential influence of a researcher's "world view" on the treatment of outlying cases. The assumptions taken for this approach include that outliers discussed here are not based on some kind of error that happened in the data collection phase (e.g. measurement error).

Most statistical tests are based on probability theory, which in frequentist terms, takes care of disconfirming cases to a certain extent. This is particularly true in a frequentist interpretation of probability, where a certain percentage of trials supports a specific argument. Talking in strict critical rationalist terms, probability confirmation would not be tolerated and thus one single case opposed to the remaining data would "crack the test procedure". As a certain level of "error" or deviation is allowed, one can conclude – irrespective of studying its underlying mechanisms – that deviant and disturbing cases would be labelled similarly or even excluded. Still, those test procedures only provide tools for researchers to test theories and the way they are applied often makes the difference between different paradigms.

Presenting "distorted results" implies that no treatment is pursued. From a logical empiricist view, this procedure seems acceptable as probability theory is applied to cover for errors and to operationalize gradual confirmation. The more confirming instances a theory receives the better for the theory's acceptance. This strategy is also acceptable from a critical rationalist perspective, which urges "severe testing" and the pursuit of rejection in case of a single negative instance, or translated into frequentist probability, more than 0% confirming cases. Critical Realism accepts that knowledge of the world is fallible and even theories that are false can still be valuable if they are closer to the truth than their competitors (e.g. Popper 1962). Thus, critical realists also support this option.

Rejecting outliers is one of the options frequently pursued. This occurs by excluding outliers from an analysis, weighting them out or deleting them completely from the dataset.

This strategy aims to present undistorted results and can be seen particularly in the light of verificationism, which is frequently replaced by confirmationism and justificationism in (logical) empiricist thought (Smith 1998). Excluding or deleting cases from a sample because they are disconfirmatory with an empirical test would be counter the logic of critical rationalist thought, which seeks disconfirmation for theoretical advancement. Critical realists interested in the mechanisms that trigger the world would most likely reject the deletion of non-erroneous observations.

Accommodating outliers can occur in different ways. One very popular solution is transforming data or winsorizing (i.e. replacing extreme values by their less extreme neighbours). Even though this shapes support for theories (verificationism), it may interfere with the logical empiricist premise of accurate measurement. Similar arguments can be made from a critical rationalist perspective. Additionally, data transformation may mask the original structure of the data and thus change the severity of the empirical test. Critical realists acknowledge that reality cannot be perceived without some kind of error, although additional measures disguising or masking reality would not be advocated by them. Using robust procedures most of the time has measurement implications as, for instance, metric data are converted into ranks, which eliminates the distances between cases. Therefore, the evaluation of this kind of data treatment is very similar to the transformation of data.

Performing sensitivity analysis means calculating two results (including and excluding outliers) and presenting the difference. This approach may be highly related to a critical perspective on empirical data analysis. For the logical empiricist, sensitivity analysis could allow the researcher or the reader to interpret the result based on values or in a subjective way rather than a fact basis, which is undesirable. For the critical rationalist, conducting sensitivity analysis could be interpreted as two different scientific tests as long as they do not lead to inductive reasoning afterwards. For the critical realist supporting the “non black box mode” of statistical analysis, the influence of outliers on a statistical model can be informative and is thus welcome.

Incorporating outliers in a revised model (e.g. using mixture model) is desirable from a logical empiricist point of view if the revised model receives a higher level of empirical support as it reduces “errors”. For the critical rationalist, the rejection of the initial model is important as it helps to advance theory. The test of the changed model can be considered as just an additional empirical test and thus part of a new conjecture and refutation attempt. A critical realist researcher testing different models will consider those that support extreme and outlying cases to be superior, given that the statistical procedure used is profoundly reasoned. Nevertheless, conditions used in this modelling need to be derived from theory and not implemented in a confirmation-seeking way.

Studying outliers that do not confirm empirical regularities separately may not be the primary objective if the goal is to unveil empirical rules, such as in logical empiricist thought. From a critical rationalist point of view, deviant cases can be considered a means of advancing theory, as they offer disconfirmation. The in-depth study of those cases may help to develop new conjectures that are then tested empirically. For him, conjectures are the basis of scientific progress. In “Three views concerning human knowledge”, Popper (1956) proposes to use crucial experiments to decide between alternative theories as occurs when a new theory is proposed as a superior alternative to an older theory. The new theory is tested by applying it to cases for which it yields results that are different from what is expected by the older theory. Such cases are “crucial” in the Baconian sense that they indicate the crossroads between two or more theories, but not in the Baconian sense that any theory can be established. An in-depth study of outliers may trigger additional insights on their emergence mechanisms and can, therefore, be advocated from a critical realist perspective. This is also supported by Manicas and Secord's (1983) view of the use of statistics under a critical realist perspective.

Practices contributing in the sense of creating additional knowledge are – to some extent – already in use. For instance, Cannon and Perreault (1999) cluster analyzed their data and then compared the results with and without outliers. So did Webster (1996) for regression analysis and Raghubir and Srivastava (2002) for ANOVA. Slotegraaf, Moorman and Inman (2003) calculated the sensitivity of the results and included the extreme cases in a hierarchical nested model. Chaturvedi, Carroll, Green and Rotondo (1997) also discussed an unstable, outlying and small market segment arising from cluster analysis along with the normal segments. Even though mentioned in the statistical literature, none of the papers in the publication audit identified outliers in a separate study. Julander and Söderlund's EMAC paper (2002) should be mentioned as it identifies about one third of the cases that reacted differently to a bad service encounter – they were still satisfied, as trust acted as some kind of glue and kept people satisfied. The option of incorporating and investigating outliers separately may be quite similar in the case of a cluster analysis.

2. *Extreme cases in marketing – a publication audit*

Marketing literature has developed a substantial body of knowledge on extreme and exceptional phenomena, i.e. people standing out of the crowd in terms of their behaviour, knowledge, attitudes, etc. People exhibiting such behaviour have been labelled lead users or extreme brand-loyal customers, for example. They may also arise when surveying an "average" population and may then manifest as extreme observations and thus evoke the same issues as discussed previously.

In marketing literature, no information could be found as to whether these problems exist with empirical studies. This section of the dissertation, therefore, investigates how extreme cases are identified and handled in the analysis procedure. For that purpose, a publication audit similar to a meta analysis is conducted. While a meta analysis, (e.g. Hunter, Schmidt and Jackson 1982, Glass, McGaw and Smith 1981, Tensaout 2006) combines the results of several studies that address a set of related research hypotheses in a statistical way, a publication audit qualitatively investigates questions related to the content and procedures used in the respective literature area. In this case, categories were initially created from the statistics literature reviewed in section 0 and then adapted.

2.1. Audit scope and methodology

The following literature analysis parallels Armstrong, Brodie and Parsons' (2001) study on the use of hypotheses in marketing. Because of the rather narrow definition of extreme values and outliers, they will be called "exceptional cases" to avoid prejudice if no specification is possible. In a few articles, exceptional case treatment was pursued twice (e.g. for two different experiments within this same article). These articles were also included twice to account for differences in analysis procedures and treatments.

For the selection of the journals, the Social Sciences Citation Index (SSCI) from the year 2006 has been used. The four premier Marketing journals have been selected and, in order to receive indications as to whether practices differ in less prestigious journals, a "second tier" journal is taken as reference. The Journal of Marketing (JM, SSCI = 4.83), Marketing Science (MS, SSCI = 3.98), the Journal of Marketing Research (JMR, SSCI = 2.40), The Journal of Consumer Research (JCR, SSCI = 2.04) and the Journal of Advertising Research (JAR, SSCI = 0.48) are included in the analysis from January 1990 through mid-2006. Slightly different time periods shown in Table 21 relate to the different publication policies of the respective outlets.

All 3918 articles in the given time period were included in this study. The JSTOR database and the EBSCO database were used to complement each other as none of them covers the whole range of articles in full text. Duplicates were not included.

Journal	Period covered by analysis	Articles covered
Journal of Marketing	Jan 1990 - Oct 2006	920
Journal of Marketing Research	Jan 1990 - Nov 2006	929
Journal of Consumer Research	Jan 1990 - Dec 2005	693
Marketing Science	Jan 1990 – Fall 2005	473
Journal of Advertising Research	Jan 1990 – Mar 2006	903
	TOTAL:	3918

Table 21: Articles covered by analysis

First, a list of search terms from statistics and marketing literature was composed. Then, the journal databases were scanned for these nine terms. The results of this database query can be seen from the "hit" list shown in Table 22. The table provides an overview on the relative importance of each of the search terms and combinations of them (e.g. influential and observation).

Term	Frequency of appearance
Outlier	149
Contaminant	12
Influential (observation, case, value)	3
Fringelier	0
Extreme (observation, case, value)	212

Table 22: Outlier terms used in literature

First, publications containing the keywords were investigated for relevance. The coder determined whether the term used was related to an empirical study. Particularly the term "extreme" appeared more frequently in non-empirical sections of the papers. In the event that one of the terms appeared within the empirical study of a paper, the empirical study was investigated. The relevant text sections were copied to a separate document and analysed. Next, the analysis procedure was recorded as it was mentioned in the paper. Then, the coder noted whether an exceptional case identification procedure was performed and whether a description of these extreme cases was given. The type of outlier management and the reasoning for a particular approach concluded the coding tables. In total, 363 papers were completely coded. The difference to the total frequency of appearance of 376 accounts for multiple nominations. Finally, 109 papers dealing with the "outlier phenomenon" in data analysis, and not merely mentioning the search term, were entered into the analysis.

2.2. Audit results

In the time period under investigation, the Journal of Marketing revealed 27 relevant papers; the Journal of Marketing Research contained 35 publications, the Journal of Consumer Research 15, Marketing Science 17, and the Journal of Advertising Research 15 relevant publications. The percentages of outlier treatments mentioned in the publications are displayed in Table 23. These ratios were corrected with the help of data on the frequency of empirical studies in the respective journals provided by Hubbard and Lindsay (2002). Assuming that the analysis procedures concerned are almost evenly distributed across journals, the corrected ratio numbers give an indication of the importance attributed by the marketing community.

Journal	Relevant articles/ total articles	Empirical studies	Corrected ratios
Journal of Marketing	3,04%	74,30%	4,10%
Journal of Marketing Research	4,31%	90,90%	4,74%
Journal of Consumer Research	8,02%	81,40%	9,85%
Marketing Science	7,08%	74,80%	9,47%
Journal of Advertising Research	1,66%	80,50%	2,06%

Table 23: Relevant articles per journal

Analysis methods

Because of the small numbers of relevant articles, per journal reporting has been waived in favour of a summated description. The categorization of analysis methods follows the similarities in outlier treatment as well as the historical distinction between correlation and experimental research in psychological literature (Cronbach 1957, Field 2000). This categorization is extended with more recently developed procedures. We acknowledge that some of these procedures, even though mentioned in different categories, may be rather similar to each other from an application point of view (e.g. Cohen 1968) (see Table 24).

Category one includes techniques based on factor analysis and approaches similar to structural equation modelling and partial least square modelling. The second category covers analyses closely linked to measures of central tendency and variation, such as analyses of variance, co-variance and chi² techniques. The third category includes linear regression modelling, hierarchical and correlation models. Category four contains logistic regressions and logit models. The fifth category covers classification methods such as clustering and latent class modelling. Category six involves various types of econometric investigations. Because of their relatively low frequency, all other tools such as various types of matrix algebra have been classified as "other".

Analysis method	Count	Percentage
Factor Analysis, SEM and PLS	12	10,43%
Central tendency / Analysis of Variance (Anova, Ancova, Chi²,…)	37	32,17%
Correlational analyses (Regression, hierarchical modelling,…)	39	33,91%
Logistic Regression, Logit Analysis	10	8,70%
Classification Methods (Clustering, latent class)	9	7,83%
Econometric Modelling	4	3,48%
Other	4	3,48%
TOTAL	115	100,00%

Table 24: Analysis methods

Identification and description

Procedures for the identification of exceptional cases are closely related to the method of analysis. However, it is important to know which procedure leads to the decision for the subsequent outlier management. Whereas most of the articles did not specify the exact procedure used for identification, 78.26% (90 of 115) noted that some kind of identification step was pursued. About 36.52% (42/115) described at least the size and other relevant characteristics of outliers.

Management and reasoning

In the papers, a total of six categories of outlier management could be identified. Some authors, especially those confronted with less extreme and disturbing outliers, used an inclusion approach. Other authors calculated with and without outliers and reported both results. Few authors chose analysis procedures insensitive to more extreme values. Another strategy – although very much linked to the use of robust procedures – is the transformation of values or the weighting of variables. The most frequently used approach to managing outliers consists of excluding or deleting extreme and outlying observations. Table 25 depicts various management approaches and their frequency of use in the respective time periods.

Management approach	Count	Percentage
Include cases	7	6,09%
Report results with and without outliers	17	14,78%
Use robust / outlier insensitive procedures	9	7,83%
Transform values / weight variables	15	13,04%
Exclude / Delete outliers	57	49,57%
No outlier found	5	4,35%
Not Mentioned	5	4,35%
TOTAL	115	100,00%

Table 25: Management approaches in literature

Most of the articles also argued in favour or against a particular management approach. Because of the very different reasons given, only three broad groups could be identified (Table 26). Statistical reasons include all arguments given without taking note of the research question at hand (i.e. interpreting the exceptional cases in the context of the study). Contextual reasons cover all types of arguments in relationship with the purpose of the study such as which and why certain values are not realistic for a particular variable.

Arguments for respective management approaches	Count	Percentage
Statistical reason	57	49,57%
Contextual reason	28	24,35%
No reason given	30	26,09%
TOTAL	115	100,00%

Table 26: Arguments for outlier management

2.3. Discussion

Surprisingly, the amount of publications reporting the finding of outliers is very small (less than 10% of all journal publications under investigation). Two factors should, nevertheless, be considered when looking at these numbers: First, the file drawer problem, which is most relevant to many meta studies, may also be inherent to this study. Significant results have a much higher probability for publication than non-significant ones (which end in the researcher's file drawer). Estimates suggest that significant findings are eight times more likely to be submitted and, for example, 97% of the articles in psychology journals report significant results (Field 2000). Consequently, we suspect that a certain number of publications affected by outliers have not succeeded in being published. Second, only a proportion of the methods used in these journals are sensitive to the outlier topic. Even though the data have been corrected for empirical studies, qualitative studies for example do not label the outlier issue as such as discussed in the section on the relativist paradigm. Furthermore, the number of publications affected with outliers may be underestimated as many papers do not declare whether the assumptions for statistical tests (including outliers) have been investigated. A recent publication in Educational Psychology revealed that only 8% of the papers declare the results of assumption testing (Osborne, Christiansen et al. 2001).

An uneven distribution appeared across the four journals under study. While only 2% of the empirical studies in JAR reported extreme observations, almost 10% in the Journal of Consumer Research did so. This difference might result from a different methodological focus (some methods do not result in extreme observations), review and editorial policies as well as non-reporting of these cases. The more frequent appearance of extreme observations when using "correlation methods" and "analyses of central tendency" seem related to the method of calculation.

The identification and management section of the audit empirically underlines the indications given by various researchers in the field of marketing, psychology and business research (e.g. Julander and Söderlund 2002). Researchers tend to exclude or average out extreme observations in order not to "disturb" their result. This is especially relevant for the management of exceptional cases. *The majority of publications discussing the treatment of exceptional cases belong to the category "exclusion of outliers". Interestingly, around 15% of the publications pursued a "report with and without outlier approach", which can be considered as a more "objective" and less researcher-determined approach to treating these cases*. One paper illustrating this approach is Agrawal and Kamakura (1995). While this approach is not mentioned explicitly in the statistics literature, it can be interpreted as a test of discordancy and has been documented by e.g. High (2000).

About one half of the arguments for a specific treatment focussed on statistical reasoning. Arguably, a statistical argument allows taking the most clear-cut decisions instead of discussing these exceptional cases in depth. On the other hand, issues that are handled in very similar ways within a community, give rise to the speculation that common mental models or paradigms influence working practices. Statistical reasoning accumulates even more doubt in the light of the statistics literature arguing hat the final decision on how to treat these cases is subjective und does not provide threshold levels for exclusion purposes.

Comparing the practices displayed here with the research paradigms discussed in the previous section, one may perceive support for the logical empiricist paradigm in several aspects. For instance, excluding/deleting or weighting out/transforming values as well as statistical arguments can be attributed to a logical empiricist approach. The reporting of result sensitivity can be attributed to a critical research approach. Interpreting only the attitude towards outliers as an indicator for a specific paradigm, one recognizes the dominant role of logical empiricist practices in the area of outlier management (62.6% of papers under study) and argumentation (49.6 %). While this interpretation may be brave, a future study using multiple indicators (such as those discussed in the next section) may give this estimation a more solid base.

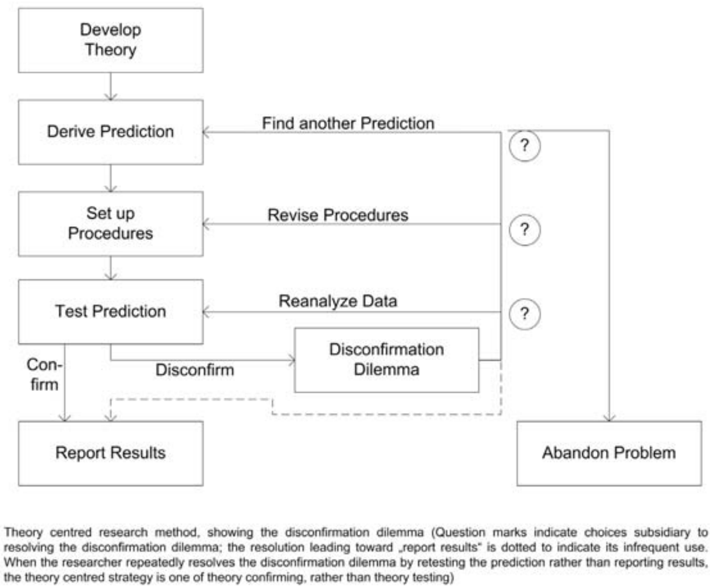
3. The "outlier problem" – standalone issue or part of a bigger phenomenon?

While public opinion frequently regards scientists as advocates for true knowledge driven by honourable motivation, empirical evidence shows that this is not necessarily always the case. As discussed in section 0, uncritical deletion or averaging out of outlying observations can deliver more "appealing" results in data analysis, such as improved test power, lower α values or higher coefficients. This consequently leads to the confirmation of hypotheses or research questions. Besides the handling of outliers, confirmation bias and the dominance of a research paradigm may also manifest in other aspects of the process. This section investigates phenomena with similar symptoms.

3.1. Criticism of current research practices

Frequent criticism focuses on the type of hypotheses used and the uncritical application of hypothesis testing. Armstrong, Brodie and Parsons's (2001) publication audit of six leading marketing journals (1984 – 1999) distinguishes three types of hypotheses: *Exploratory studies (12.6%)* depart with no formally specified hypotheses and are mostly used when one has no or only little knowledge about a phenomenon. When using *competing hypotheses (13.0%)*, the researcher investigates evidence for two or more plausible hypotheses. This approach is frequently recommended as it changes the researcher's role from advocating a single hypothesis to evaluating which of a number of competing hypotheses is best and impedes scientists from stopping with a premature conclusion. *Single or dominant hypotheses (74.4%)* are designed to rule out null hypothesis, e.g. the absence of a relationship between two concepts. This approach is particularly useful (1) after the exploratory phase to help refine a plausible hypothesis on a topic; (2) when it may not be feasible to develop competing hypotheses; (3) when it may be too costly to test alternatives; (4) when an efficient "market" for ideas exists, such as when parallel teams pursue solutions to the same problem at the same time, with well established criteria for the evaluation and good communication among teams or (5) where the task is to clarify the conditions under which an accepted hypothesis holds. The share of this group can be estimated to be even higher as in less than three percent of the cases, competing hypotheses differed substantially from one another. This is particularly problematic, as other studies have shown, as the use of a single hypothesis leads to a bias in the way people evaluate evidence (e.g. Lord, Ross and Lepper 1979). *Greenwald, Leippe et al. (1986) subsumed that the use of a single hypothesis does not effectively promote objectivity.*

Greenwald, Leippe et al. (1986) suggests that a confirmation bias exists with researchers. *Applied to hypothesis testing, this confirmation bias renders "real" theory testing approaches into "theory confirming" ones*. The latter ones repeatedly retest the predictions rather than report disconfirming results. In more detail, this means that if the predictions made from a theoretical base differ from the results of hypothesis testing procedure, the researcher faces a dilemma, also known as "disconfirmation dilemma" (Greenwald and Ronis 1981). This dilemma can be solved in the following ways (see also Figure 4 for illustration).



Source: adapted from Greenwald, Leippe et al. 1986

Figure 4: Disconfirmation dilemma

(1) The theory being tested is incorrect and, therefore, disconfirming results are reported or (2) the researcher perseveres at testing the theory's correctness by, for example, conducting additional data analyses, collecting more data or revising procedures (Greenwald, Leippe et al. 1986).

A confirmation bias exists if interpretations and judgments based on new data tend to be overly consistent with preliminary hypotheses. A series of experiments supports the claim of a confirmation bias existing among researchers. Mynatt, Doherty and Tweney (1978), for example, showed that the scientists under investigation seldom sought disconfirmation of their favoured theories; they ignored information that was falsifying their theories. Reasons for this phenomenon can be multifaceted. MacCoun (1998, p269) classifies confirmation bias under strategy-based errors, which “occur when the judge, due to ignorance or mental economy, uses “suboptimal cognitive algorithms”. A similar heuristic has been unveiled by Gigerenzer (1991), labelled “tools to theory”. He argues that methods of justification shape theoretical concepts. This means that tools, such as statistical tests and the subsequent software, are introduced and in this way suggest new theoretical metaphors and concepts. As soon as these tools are accepted by an individual, the new theoretical metaphors and concepts are also more likely to be accepted by the scientific community if their members also use this tool.

Low rejection rates of hypotheses are another indicator. Particularly low ones have been identified in social sciences (Hubbard and Armstrong 1992). Wilson, Smoke and Martin (1973) investigated three major sociological journals and found that 61 of 76 (80.3%) of the articles using significance tests rejected the null hypothesis. Greenwald’s (1975) investigation of the 1972 volume of the Journal of Personality and Social Psychology unveiled 175 of 199 (87.9%) of the articles managing to reject the null hypothesis. For the articles on budgeting and control published in three leading accounting journals, Lindsay (1994) unveiled a rate of 84.2%. Finally, Hubbard and Armstrong (1992) analyzed 692 randomly selected papers that used significance tests, 92.2% of which rejected the H_0 hypothesis. Smart’s (1964) results also suggest that a certain development takes place. He unveiled that 30% of studies in doctoral dissertations in psychology failed, while this ratio dropped to 20% of papers presented at the annual American Psychological Conference and to 10% of papers published in journals. Reasons for this process could be linked to an increasing filter of control by the academic reviewing system and the increase in knowledge during a researcher’s career. Overall, one can also assume that in the test situation, confirmation is sought.

Criticism from the statistics community even points to a “hypothesis testing ritual”. Many researchers, particularly in the field of psychology and education, engage in “mindless hypothesis testing in lieu of doing good research: measuring effects, constructing substantive theories of some depth, and developing probability models and statistical procedures suited to these theories" (Luce 1988). One main observation here is that researchers follow a null hypothesis testing ritual that most often includes the following steps: Setting up a statistical null hypothesis of "no mean difference" or "zero correlation" without specifying the predictions of the research hypothesis or other substantive hypotheses, the use of 5% as convention for rejecting the null hypothesis and, in case of significance, the acceptance. Always performing this procedure is part of the ritual (Gigerenzer, Krauss and Vitouch 2004). Moreover, null hypotheses are set in a way that they are (almost) always rejectable. Nevertheless, many studies lack the information of the pre-probability, i.e. what was the probability of rejecting the null hypothesis before the experimental treatment, and what is the probability of wrongly rejecting an hypothesis (Type II error). Thereby, the authors criticize a lack of understanding (and interest) in significance testing among researchers and the further diffusion of error-laden knowledge in order to pursue their well-established practices.

Publishing practices of academic outlets also play a certain role in this area. Armstrong (2007) extends this view with the argument that even journal reviewers misinterpret statistical significance (Atkinson, Furlong and Wampold 1982). After closely reviewing papers using significance tests, he concludes that their use should be abandoned as they harm scientific progress. The main target of this critique lies in the unique focus on p-values and incorrectly using them to make dichotomous decision as to the truth of a null hypothesis (i.e. a parameter equals zero in a population, which is almost always false). For their 1991 article, Armstrong and Hubbard administered a mail survey with the editors of 20 leading psychology journals. "For a sample representing 32 journal-years, we were able to locate only one paper with controversial results that was unanimously recommended for publication by the reviewers. I expect that this situation also exists in marketing journals." (Armstrong, Brodie et al. 2001, p72). In his 1977 study, Mahoney tested unsuspecting reviewers with two versions of a paper: one supporting conventional and one supporting controversial findings. While those receiving the conventional version accepted it, the controversial paper was broadly rejected. Therefore, one may also conclude that the "reviewer filter" promotes confirmation bias.

Dunbar's (1995) study gives insight in to the processes in "real world laboratories". Scientists in the four laboratories under study were quick to modify their hypotheses when they received inconsistent evidence. Even if evidence called for change, they tended to be resistant when working alone. In some cases, researchers even altered data to support their hypotheses (Armstrong, Brodie et al. 2001) or selectively deleted studies from a meta-analysis by investigating their preferred hypothesis.

MacCoun (1998) contributes to the discussion with a categorization of biased evidence interpretation by researchers. He starts from a purely researcher-related perspective and identifies three factors: intentionality (a combination of consciousness and controllability), motivation (based in the judges preferences, goals or values) and normative justification (justification is always relative to a system of norms). In his opinion, a continuum exists from 'cold bias', an unintentional and unconscious form that occurs even when the judge seriously seeks for accuracy, and the "hot bias", which is unintentional and perhaps conscious, but directionally motivated – the judge wants a certain outcome to prevail.

The confirmation bias may not be a problem, if id would not to have serious consequences. Using the 'sleeper effect' as an example, Greenwald, Leippe et al. (1986) show that confirmation bias can lead to a significant delay in scientific progress - in that particular case, 25 years delay to today's understanding of the phenomenon. Other examples include Wyatt and Campbell (1951) or Bruner and Potter (1964). As no systematic empirical comparison of the evolved confirming procedures with earlier disconfirming ones is attempted, the researcher is unlikely to detect the confirmation's dependence on these details, such as certain boundary conditions for the results. Consequently, over generalized results can arise and may not even be discovered as such. The discovery of these conditions is very relevant as many research designs are incomplete because of the lack of possible control over all aspects. Furthermore, replications and extensions are very rare even though they are claimed to be the key to generalizations (Leone and Schultz 1980). The generalization across all contextual variables is inevitably in error. Oftentimes, there are conditions among the infinity of unexamined contextual variations, under which the results do not hold. These could probably be drawn from competing, more developed theories.

What are the reasons that researchers continue to use dominant hypotheses, pursue the null hypothesis rituals and take into account confirmation bias even though knowing better? In their study on accounting researchers, Borokowski and Welsh (2000) revealed that approximately 70% of the researchers were aware of the existence of a confirmation bias. The answers given by many researchers as to why they tolerate confirmation biased procedures can be grouped into three areas:

First, confirmation bias can be a valuable heuristic if no other superior alternative is available or known (e.g. MacCoun 1998). Secondly, experiencing success by having predictions confirmed shapes the researchers' expectations of future success and, if researchers are confirmation biased, they continue their activity in the same, biased way. Thirdly, confirmation bias strengthens the establishment that practices it. This can be seen as a conservative force maintaining the status quo. A veteran, confirmation-biased researcher, for instance, might be passing on his knowledge to the researchers of the next generation and, if in the same school of theory, there are "tangible rewards of status and resources for perpetuating the theory via collective, confirmation biased methods" (Greenwald, Leippe et al. 1986, p223). This can also be one reason why the approach of competing hypotheses, already discussed at the end of the 19^{th} century (Chamberlin 1890) and considered to progress science faster, did not succeed. In economics, Gans and Shepard (1994) found that many famous economists considered journals unreceptive to their most important papers. Re-submission experiments, where accepted publications were resubmitted after a certain time, shed a bad light on academic journals. Most of the articles were not recognized; a high number invited for revision and some even desk rejected.

3.2. Possible countermeasures

A substantial amount of measures has been proposed to overcome the shortcomings in the research process.

In terms of individual research practices, literature stresses the use of competing hypotheses, instead of single (dominant) ones, to reduce problems of confirmation bias tendency (Armstrong, Brodie et al. 2001, Gigerenzer, Krauss et al. 2004). Gigerenzer and Hoffrage 1995), for example, tested the predictions of six cognitive strategies in problem solving. The use of competing hypotheses can increase objectivity, either through the discovery of the correct hypothesis (Farris and Revlin 1989) or the general phenomenon that more disconfirming information increases the likelihood to discover correct explanations (Gorman and Gorman 1984). McKenzie (1998) found that using several alternative hypotheses, researchers were better able to judge how evidence related to each of them. Platt (1964) argued that personal attachment to our hypotheses "clouds our judgment and sets science up as a conflict among scientists, rather than among ideas" (in MacCoun 1998, p276), whereas rapidly advancing research programs would avoid these self-confirming tendencies. Besides the use of competing hypotheses or exploratory hypotheses (Brodie and Danaher 2000), active condition seeking before making generalizations has been very much advocated (Uncles and Wright 2004). Greenwald, Leippe et al. (1986) support the idea of this strategy, but rather recommend active condition seeking "in which a researcher deliberately attempts to discover which of the many conditions that were confounded together in procedures that have obtained a finding, are indeed necessary or sufficient"(MacGuire 1983, p223). As many researchers and critics note, this strategy is to a larger extent data driven and might lead to some even more specific case findings undermining the development of more general theories (e.g. Greenberg 1981). Nevertheless, as argued by Greenwald and Pratkanis (1988), results-centred research – as he called his approach – will end in "findings with greater shelf life than theory-centred research findings and ultimately provide the grist for better theory formulation."

One area of suggestions to overcome these biases focuses on replications. "Replication in science refers to repeating a research study to discern the reliability and validity of a set of findings" (Harlow 1997, p8). Thereby, three kinds of replications can be distinguished: In literal replication, a study is repeated exactly as the original; in operational replication, the attempt is made to reconstruct the main procedures in a study; constructive replication means that "another study of the same constructs is conducted, possibly using different sampling, procedures and measures" (Harlow 1997). Meta analysis helps to summarize a set of findings of one particular phenomenon identified across a wider variety of studies.

Armstrong (2003) argues that important findings in other disciplines are frequently characterized by replicability, validity, usefulness and a "surprise factor". However, this does not seem very relevant to the marketing area: Applying these criteria to marketing publications, only a small number of publications accounts for an "important finding" (Armstrong 2003). Some of these criteria probably relate more to one specific philosophical view than another. Nevertheless, representatives of other "camps of thought" raised similar concerns such as Hunt's (1994) argument that marketing research tends do address "micro issues", which can be seen closely linked with Armstrong's replicability and usefulness. In their 1994 study, Hubbard and Armstrong did not find any replications at all and extensions only covering 1.8% of the 1,120 papers investigated. On the contrary, replications were considered replaced by high statistical significance in the original study (Oakes 1986) and less relevant in editor's decision for publication (Kerr, Tolliver and Petree 1977).

Minimizing error in the data collection context includes the use of "less noisy" methods (e.g. paired comparison instead of rating), a proper experimental environment (e.g. individual testing instead of large classrooms), proper motivations (by performance contingent payments instead of flat sums), unambiguous instructions and avoidance of unnecessary deception of participants about the purpose of the experiment, which can lead to second-guessing and increase variability (Gigerenzer, Krauss et al. 2004).

Instead of completely banning significance testing (Armstrong 2007) from social science research, earlier literature (e.g. Harlow 1997) takes a less extreme stance. The following points can be summarized from the literature:

Many researchers reject the mechanical use of any method and seek to replace it with reflected judgement. Tukey (1969), for instance, promotes "high standards of statistical reasoning, but without specifying a single model of statistics which might serve as a criterion or quality of reasoning" (p90). Kirk (1996) emphasizes the concept of practical significance of data, which cannot be linked to a ritual use of a single procedure. Finally, Thompson (1996) interprets researchers' constrained use of mathematical calculation of probabilities as "a purely atavistic escape … from the existential human responsibility for making value judgements…[however,] empirical science is inescapably a subjective business" (p28). Cohen (1994) reminds that a broad array of statistical techniques is available to be used by the researcher, but they must be chosen sensibly and by informed judgement.

Using an enlarged statistical toolbox is frequently recommended, particularly by the statistics community. Confidence intervals are widely used in many scientific disciplines (see e.g. Schmidt and Hunter 1997). They estimate the value of the parameter and provide a band of confidence with upper and lower limits. If the parameter values of the null hypothesis do not fall within the confidence interval, the null hypothesis can be rejected at the corresponding alpha level. Confidence intervals, therefore, provide more information than conventional null hypothesis testing procedures (Cohen 1994, Harlow 1997).

The reporting of effect sizes and power can already improve transparency (Hubbard and Armstrong 1997). Effects sizes include the magnitude of difference between hypothesized and sample value such as means or proportions as well as squared measures of the strength of association (e.g. r^2) (Harlow 1997). Power relates to the probability of correctly rejecting the null hypothesis. It is treated in more detail in Exhibit 2 (Section. 0).

Researchers should evaluate how well a model approximates the data, but not necessarily focus on statistical significance. Structural equation modelling has become increasingly popular in the last years and is particularly aimed at investigating the fit of the data to the theoretical model. Fit indices are used to measure the degree of approximation. “In SEM we are essentially trying to retain the null hypothesis that our model offers a good fit to the data” (Harlow 1997, p7).

Null hypothesis testing has a long history in social science. If used to make dichotomous decisions for well-reasoned and specific hypotheses – and when supplemented with additional scientific input such as effect sizes, power, confidence intervals and sound judgement – it can be very effective in highlighting hypotheses that are worth further investigation (Harlow 1997). However, many voices do not explicitly subscribe to this view and propose alternative methods (e.g. Cohen 1994). Currently, many null hypotheses are used as suggested by many software programs: No effects, while alternative hypotheses suggest some non-zero effect. Many authors currently advocate the use of specific, defeatable hypotheses as well as non-zero values (as some minimum correlation is given among many variables in social science datasets) (Harlow 1997). To address the “Null hypothesis ritual”, leading authors (e.g. Cohen 1994, Gigerenzer, Krauss et al. 2004) recommend using an enlarged statistical toolbox adapted to the relevant situation. John Tukey’s (1977) exploratory data analysis seeks not to generalize on a population level but proposes the use of simple, graphic based and informal techniques, which were further enhanced by Tufte (1983, 1990), Cleveland (1993, Cleveland and McGill 1988) and Wilkinson (1999).

"Bayesian statistical methods support inferences…[through the] use of prior information and empirical data to generate posterior distributions, which in turn serve as the basis for statistical inferences." (Pruzek 1997, p287). These methods are often referred to as subjective as they allow the inclusion of a researcher's previous experience and beliefs in the current analysis. However, they also allow for the account of the "base rate" (e.g. the effect which is already present before the experimental treatment) and, therefore, achieve a less overestimated estimation of effect size (Harlow 1997). The acknowledgement of existing limitations in many traditional approaches is not only limited to social science, but to many scientific disciplines. In his recent article published in Science, Malakoff (1999) even speaks about a contemporary Bayesian revolution.

Another domain of recommendations deals with the review process. For a more detailed overview, the interested reader is referred to existing literature (e.g. Armstrong 2003, MacCoun 1998).

Overall, one may assume that a confirmation bias exists in social science literature and consequently also in marketing literature. In many areas, awareness and counteractions exist such as, for example, of the "null hypothesis ritual". Some measures have been outlined in section 1.3. Compared to this stream, the outliers issue has not received as much attention.

D. Empirical Study

This chapter studies the effects of outlier treatment on the outcome of an empirical study. As discussed previously, the study of outliers is closely linked to the study of their sources and, in particular, whether they are real variations or errors. Thus, a well-researched domain can support the research process as other studies can serve for validation. Tourism in Australia is ideal with respect to this criterion as the Tourism industry is well developed. Data are collected in many different areas and a wealth of academic research is available. For the conceptual base, market segmentation is chosen because of its popularity and the availability of empirical data. To minimize influences from the data collection instrument and the recording situation, which could result in errors, analysing past behaviour is advocated as it allows for cross checks through, for example, proofs of purchase such as tickets.

The present study investigates outliers separately and applies a sensitivity analysis. These strategies presented in section C.1.3 are compatible to a critical realist approach. In terms of statistical method, a cluster analysis procedure is used. To test the effect on the results, censoring rules similar to those used in the published reference study are applied and the results before and after are compared. Furthermore, outliers in this dataset are investigated and probed as to whether they can help to derive additional insights. As a reference, the segmentation study on international tourists to New Zealand by Becken and Gnoth (2004) is used. Similar data for Australia are available from the Bureau of Tourism Research (BTR). Australia and New Zealand are also comparable in their tourism industries. For instance, their main "markets" for tourists in Europe are the same: Germany and the UK. In both Australia and New Zealand, visitors predominantly arrive by airplane. Finally, similar tourist types have been identified in separate studies, such as backpackers, group tourists and campervan tourists. Australia and New Zealand can also be compared in terms of their marketing activities and strategies (see e.g. Riege and Perry 2000). Furthermore, the analysis is based on behavioural data, which can be recorded in a more reliable way than, for example, psychographics.

In short, this section aims to find answers to the following research questions:

Does a priori elimination of outliers and extreme cases through, for example, sample censoring or pre-segmentation yield different results compared to calculation with the entire data?

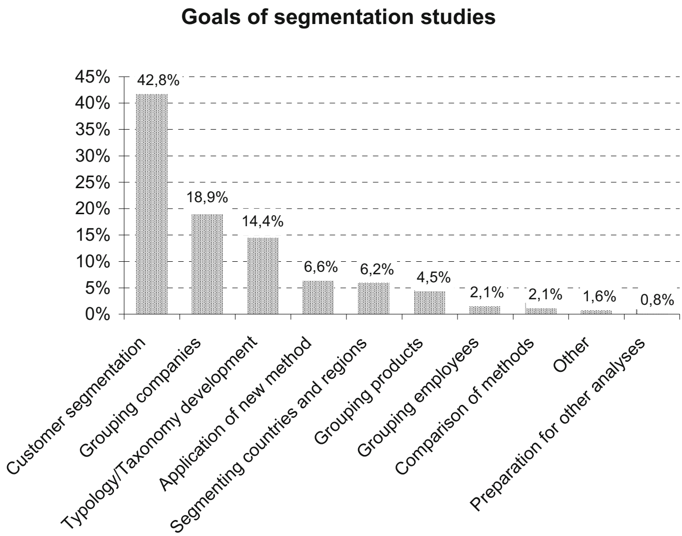
Does studying outliers and extreme cases result in additional theoretical insights?

This section's contribution also includes proposing and testing a method to analyse groups of extreme behaviour and accounts for the frequent demand for "replication and extension studies" related to scientific progress in general (e.g. Armstrong 2003) or under a critical realist view in particular (e.g. Robson 2002).

1. Literature review and theoretical background

As this study replicates a segmentation study, the background of this concept is reviewed first generally and then specifically for the field of tourism. Segmentation as a management technique aims to group people or organizations to achieve a higher degree of group internal homogeneity while obtaining heterogeneity between segments. This technique is widely used and rather broad as a concept. Some of the most widespread segmentation approaches include normative segmentation, niche marketing, micro marketing, database marketing, relationship marketing and mass customization. Each of these conceptualizations includes different aspects and popular practices (e.g. Kara and Kaynak 1997). Overall, segmentation is often linked to long-term strategic decisions within an organization or for a destination. Baumann (2000) conducted a meta-analysis on market segmentation and identified 247 studies from the beginning of the 1970s to April 2000 by studying 83 academic journals including all prestigious marketing journals. Her analysis and findings serve as a "benchmark" for the present study, allowing generalizations within stated limits (Easterby-Smith and Thorpe 2002).

Segmentation studies are conducted for different reasons as shown in Figure 5: Customer segmentation relates to individuals who are grouped in order to, for example, identify target market(s) and apply marketing measures in a more precise way. Grouping companies applies to customer companies as well as the identification of industries and strategic groups and represents the second largest group of studies. Finally, the development of typologies and taxonomies reaches the third rank with about one seventh of the studies.



Source: adapted from Baumann 2000

Figure 5: Goals of segmentation studies

Theoretical foundations for segmentation studies have been very diverse and depend on the subjects used as the basis for segmentation. Some authors, for instance, used the involvement construct as a basis to segment consumers, and was sometimes linked with the discussion of cultural differences in consumption (e.g. Amine 1993, Assael 1987, Wills, Samli and Jacobs 1991), while others used widespread micro and macro economic concepts (Frank, Massy and Wind 1972).

Customer segmentation as a business strategy aims to manage diverse customer needs by offering specific products to customer groups with similar preferences. This concept has been widely discussed in the academic literature, applied in companies, and found its way in to almost every marketing textbook in use (e.g. Dubois, Jolibert and Mühlbacher 2007, Kotler and Keller 2006). Six critical factors for successful market segmentation have been outlined in the literature (Frank, Massy et al. 1972). They are displayed in Table 27.

Criterion	Relevance
Identifiability	Criteria that serve for differentiation between segments should allow a clear assignment of cases to a segment. Easy measurement of those criteria facilitates application in business.
Substantiality	Costs of addressing the segment should be in line with expected benefits by the company. Segment size plays a role here.
Accessibility	Segments should be reachable, particularly in terms of distributional and communication measures.
Stability	Segments should be stable (composition, behaviour, etc.) over a certain period of time, otherwise segment specific measures may not target properly.
Homogeneity in behaviour	Cases within a segment should behave similarly and distinct to other segments.
Segment handling	Segment specific information must be usable for targeting the segment appropriately. Capabilities and goals of the company should allow the satisfaction of those goals.
Source: Frank, Massy et al. 1972, Steger 2007	

Table 27: Critical factors in market segmentation

In the process of conducting market segmentation, criteria need to be determined and measured. To date, the classification of these criteria as shown in Table 28 by Frank, Massy and Wind (1972) is still considered very relevant (Wedel and Kamakura 2000).

	General criteria	Product specific criteria
Observable	Cultural, geographical, demographic	Product expertise, frequency of use, product loyalty, situation
Not observable	Psychographics, values, personality, lifestyle	Psychographics, benefit, perception, elasticity, attributes, preferences, intentions
Source: Frank, Massy et al. 1972, Steger 2007		

Table 28: Classification of segmentation criteria

Overall, a wealth of criteria is available to the marketer across the four cells, with different advantages and disadvantages. Data for general, observable criteria are easy to collect, while measures related to needs are difficult to derive. The opposite is true for product specific, non-observable criteria that are difficult to collect, but very relevant for purchase decision making (Bonoma and Shapiro 1984). Therefore, criteria are frequently combined with a focus on using a few highly explanatory ones. The most effective ones can be found in the group of product specific, non-observable criteria (Wedel, Kamakura and Böckenholt 2000). The reference study used in this chapter applies observable, product specific criteria such as transport or accommodation choices.

After the decision for segmentation criteria, the segmentation method needs to be determined. A wealth of methods has been proposed in the literature that can be distinguished into descriptive vs. predictive and a priori vs. post hoc methods (see Table 29).

	A – priori	Post hoc methods
Descriptive	Crosstabulation, log - linear modelling	MDS, neural networks, hierarchical clustering, latent class analysis, hierarchical Bayes modelling
Predictive	Regression modelling, logistic regression, Discriminant analysis, Structural equation modelling	Neural networks, Decision trees, latent class analysis, hierarchical Bayes modelling
Source: Wedel and Kamakura 2000, Steger 2007		

Table 29: Segmentation methods

A priori methods determine the number of segments and their separation criteria before the analysis is pursued. Many applications can be found for usage frequency (e.g. heavy vs. light users). More generally, this application makes sense if information is needed for assumed or existing groups. They are often refined with additional criteria.

In post hoc segmentation, customers are assigned to groups according to a wealth of criteria. Number and segment profiles are not known before (Neal 2000). Within this group, model-based procedures pose distributional assumptions, whereas heuristic procedures refrain from doing so. Model based segmentation approaches use probability models assigning each customer a probability of membership to each cluster (Wedel, Kamakura et al. 2000). While the literature promotes model-based procedures, industrial applications prefer heuristics because they are more straightforward and less complex to apply (Hahn 2002). In terms of the general segmentation literature, the reader is referred to Steger's comprehensive overview (Steger 2007). The study used in this dissertation relates to descriptive, post-hoc segmentation using existing data and a hierarchical cluster analysis.

Customer segmentation also represents a major part of the segmentation studies conducted in Tourism literature. Dolnicar (2004) reviews the Journal of Travel Research as the most prominent outlet for segmentation studies and concludes that four types of segmentation studies exist in the literature: (1) pure commonsense segmentation, (2) purely data-driven segmentation, (3) combination of both where typically one commonsense segment is chosen and further split up into data-driven subgroups, and (4) a sequence of two commonsense segmentations. Table 30 outlines the literature identified by Dolnicar and gives a short description. Even though this overview is limited to 15 years of the Journal of Travel Research, it can be considered a representative overview as this journal represents the major outlet in the field (Dolnicar 2004).

Studying these pieces in more detail, one notes a wide array of techniques and criteria used as in general segmentation literature. However, non-observable, general criteria seem to be underrepresented.

Market segments in tourism literature have also been named "travel styles" and studied separately. Such publications include e.g. coach tourism (Oppermann 1994), backpacker tourism (Ateljevic and Doorne 2001), package tourism (Quiroga 1990, Thomson and Pearce 1980) and campervan tourism (Gnoth 1999).

Segmentation studies in tourism literature	
Pure commonsense segmentation	
Baloglu and McCleary 1999	Investigation of the difference of visitors and non-visitors of a certain destination related to the image of this tourism region.
Goldsmith and Litvin 1999	Contrasting heavy users versus light users.
Kashyap and Bojanic 2000	Explore systematic differences between business and leisure tourists with respect to value, quality and price perceptions.
Smith and MacKay 2001	The impact of age differences in pictorial memory performance for advertising message targeting.
Israeli 2002	Profiling the perception of destinations from the perspective of disabled versus non-disabled visitors.
Klemm 2002	Investigation of one specific ethnic minority in the UK and description of their vacation behaviour and interest.
McKercher 2001	Exploring systematic differences between tourists who spend their main vacation at a destination versus tourists who only travel only through that town.
Meric and Hunt 1998	Profiling the ecotourist.
Arimond and Lethlean 1996	Grouping visitors to a campground according to the kind of site rental taken and investigate the differences.
Court and Luption 1997	Grouping tourists initially by their intention to visit a destination and then searching for significant differences between those commonsense groups.
Pure data-based segmentation	
Bieger and Laesser 2002	They identify data-driven segments among the entire Swiss population.
Combination of commonsense segmentation and pure data-based segmentation	
Silverberg, Backman and Backman 1996	Nature based tourists are grouped along the benefits sought by this group.
Dodd and Bigotte 1997	Special interest group of winery visitors are segmented on the basis of demographic profiles.
Forminca and Uysal 1998	Visitors of a cultural-historical event in Italy were chosen for subsequent motivation-based segmentation.
Kastenholz, Davis and Paul 1999	They concentrate on visitors of rural areas, which are then segmented with benefit measures.
Moscardo, Pearce, Morrison, Green and O'Leary 2000	They select a (commonsense) group of visitors of local friend and relatives and then conduct a data-driven investigation.
Hsu and Lee 2002	They focus on senior motor coach travellers first and then use 55 selection attributes for further grouping.
Other / mixed approaches	
Field 1999	Domestic vs. foreign students within the segment of student travellers are explored.
Horneman, Carter, Wei and Ruys 2002	Focus on senior travellers first and then development of six segments by questioning for most preferred holiday choice.
Source: compiled from Dolnicar 2004	

Table 30: Segmentation studies in the Journal of Travel Research

Becken and Gnoth's (2004) study identified travel styles in a comprehensive, data-driven way for New Zealand, using the "tourist consumption systems" framework as a theoretical foundation. Exhibit 7 investigates this framework and its roots in consumer behaviour. The authors used a sample of travellers to New Zealand from the international visitor survey (IVS) in 2000 and segmented them related to their consumption system profile. Consumption systems include "thoughts, decisions, and behaviours regarding one activity that influence the thoughts, decisions and behaviours for a number of other activities, implying that behavioural patterns should be visible in the consumption of tourism offers" (Woodside and Dubelaar 2002, p120). This framework allows a better understanding of the whole consumption process of tourists at a destination, which is, most of the time, far more complex than the consumption patterns in other areas. Technically, clustering variables are used that cover accommodation, activity and transport choices to derive five segments. In a second and third step, discriminating variables are identified and tested as to whether they can identify the same tourist types in a different dataset. For the present study, only the segmentation procedure in the first step is used for replication and comparison. Both the availability of similar data, but also the use of censored data, make this study ideal for the investigation of the two research propositions.

As with many other segmentation studies in tourism (Dolnicar 2004), Becken and Gnoth (2004, p377) also pre-segmented their sample...

> *"into long-term tourists (those staying longer than 180 days), gateway-only tourists (those who never leave their gateway of arrival), cruise ship tourists, and touring tourists. The latter are tourists who stayed at least at two different locations within New Zealand."*

Their pre-segmentation resulted in an elimination of 23% of the cases compared to the entire sample. Taking a closer, somewhat "speculative", look at this pre-segmentation, one notes that this pre-segmentation may also influence the cluster solution presented afterwards:

Gateway only tourists stay 0 nights and likely do not use any means of transport nor pursue activities, which would result in scoring very low (or even 0) on those cluster variables. If variables are normalized by the length of stay as pursued in the study, a division by 0 would occur, which is mathematically problematic.

Long-term tourists are very likely to exhibit different behaviours in terms of transport and accommodation choices. A working holiday visa tourist, for example, may find his accommodation provided by the company or farm for which he/she is working. Transport patterns may also be different, such as staying longer at certain places and using different means of transport because of a reduced time pressure.

Cruise ship tourists very often only sleep aboard and use the ship as a means of transport, which results in no other variables (neither transport nor accommodation) being relevant.

In total, one may hypothesize that the exclusion of those diverse profiles may also impact the structure of the dataset and create a more homogeneous group of cases submitted to segmentation afterwards.

The consumption systems framework

The theoretical foundations of this framework can be viewed in different ways: Symbolic interactionist perspective as well as consumer decision making based on the theory of planned action.

In a symbolic interactionist paradigm, consumers are conceived as pragmatic actors who make sense of their worlds in terms of social realities from which they derive their identities (Mead 1934, Hogg and Mitchell 1997). McCracken (1992, p XI) postulates that consumers "use the meaning of consumer goods to express cultural categories and principles, cultivate ideas and sustain lifestyles, construct notions of self and create social change" . For the interactionist, society consists of organized and patterned interactions among individuals that can easily be observed. One can, therefore, conclude that people may travel for the same basic motivations (such as for leisure or to visit friends and relatives), but consume different services related to that travel in order to cultivate their selves.

As research ignoring product interdependencies often resulted in poor prediction of brand and product choice behaviour (Wind 1977), Douglas & Isherwood (1979), question whether a single product or brand can be seen as representative of a consumer's self concept. Therefore, different approaches of "joint consumption" have been developed. "Joint consumption" involves the consumption of two or more products or services related to each other. While Solomon and Englis' (1994, p67) concept of consumption constellations covers "a cluster of complementary products, specific brands and/or consumption activities used to construct, signify and/or perform a social role", Woodside and Dubelaar's (2002, p120) consumption systems go beyond this idea: They define their consumption system as "thoughts, decisions, and behaviours regarding one activity that influence the thoughts, decisions and behaviours for a number of other activities". Nevertheless, not all product categories seem to be relevant for the occurrence of consumption systems. Leigh and Gabel (1992) mention the connection to one's ego, public consumption and the degree of complexity as key characteristics of product categories as relevant for the symbolic interactionist view. This can be said for many aspects of tourism and travel consumption, such as accommodation, transport and activity choices. The field of leisure travel has been investigated under the perspective of symbolic interactionism several times (see e.g. Davidson and Layder 1994).

Symbolic interactionist research is mostly characterised by qualitative and interpretative methods of data collection, unveiling the meaning of a product, an activity or a service. Under this perspective, the quantitative approach taken by Woodside and Dubelaar (2002) and Becken & Gnoth (2004) would appear exceptional and quasi unique. But consumption systems also make sense from another perspective: Woodside & Dubelaar (2000) postulate that "thoughts" are linked to certain consumption processes. In their framework, "thoughts" subsumes intentions to (re-) visit, intentions to recommend a destination, visitor motivations as well as consumption related evaluations (see Figure 6).

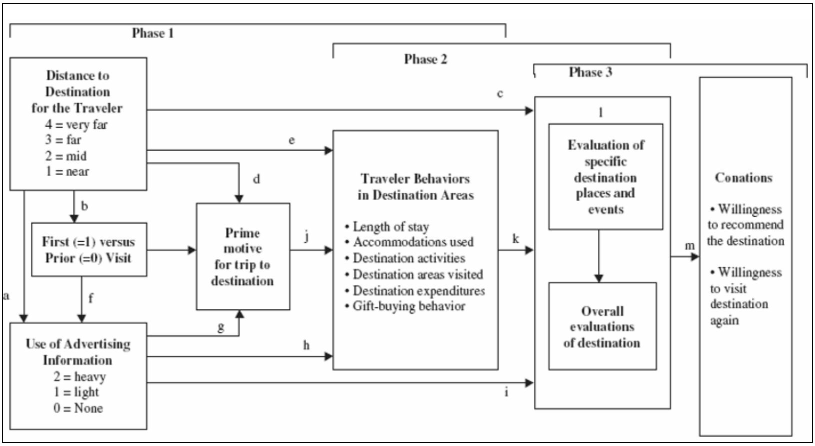


Figure 6: Tourism Consumption Systems (Woodside and Dubelaar 2002)

In consumer and organisational behaviour research, the relationship between intentions and actual behaviour has been intensively studied, mostly on the basis of the theories of reasoned action (Ajzen and Fishbein 1980, Fishbein and Ajzen 1975) and planned behaviour (Ajzen 1988, 1991, Bagozzi, Baumgartner and Yi 1992).

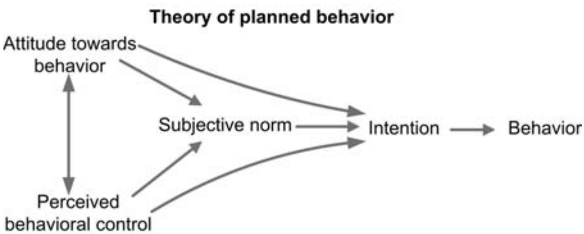


Figure 7: Theory of planned behaviour (Ajzen and Driver 1992)

Intention, as immediate antecedent for any behaviour, is preceded by the attitude towards behaviour and subjective norms including the perceived social pressure to perform or not perform certain behaviour. Figure 7 displays the main elements of the theory of planned behaviour.

The theory of planned behaviour can be considered as an extension of the theory of reasoned action that includes measures of control belief and perceived behavioural control (Armitage and Conner 2001). If individuals think that they possess more resources and opportunities and anticipate fewer obstacles or impediments, then they should perceive greater control over the behaviour. Beliefs about resources and opportunities may be viewed as underlying perceived behavioural control. While many prominent consumer behaviour models treat intentions as direct antecedent of behaviour (e.g. Peter, Olson and Grunert 1999; Engel, Blackwell and Miniard 1994; Howard and Sheth 1969), a multitude of factors and situational constraints have been identified as to why individuals deviate from their plans (e.g. Belk 1974, 1975). These variables and factors include shopping party, personality traits, proclivity to visit stores as well as compulsive types of unplanned purchases as discussed in the psychology literature (see e.g. March and Woodside 2005 p115 for an overview). Similar findings can be identified in organisational decision making: Mintzberg (1978) for instance refuses to limit "research to the study of perceptions of what those who, it is believed make strategy intend to do. And that kind of research – of intentions devoid of behaviour simply is not very interesting or productive". In a recent meta-analysis by Armitage and Conner (2001), the theory of planned behaviour was able to account for 27% and 39% of the variance in behaviour and intention respectively.

The consumption system approach, demonstrated here, includes additional antecedents such as prior product experience or advertising exposure which have been successfully tested for their influence on consumer behaviour (e.g. Hoch and Ha 1986, Holbrook and Batra 1991). Their influence on consumer behaviour has been tested in correlational and experimental frameworks. In tourism research, frameworks are predominantly quantitatively tested (see e.g. Bieger and Laesser 2002).

In the field of travel research, antecedents to intentions (and behaviour) include the distance to the destination, which often also reflects the amount of risk taken by the traveller (e.g. Fesenmaier and Johnson 1989), the influence of advertising and other information sources (e.g. Fodness and Murray 1998, 1999) and previous experience. Motivations influencing the intention to behaviour and, consequently, action are often operationalized as business, leisure or work–related.

Exhibit 7: Theoretical views on tourists' consumption systems

2. *Method*

To answer the research questions, this section addresses the effect of censoring and investigates "global" and "local" outliers. While the former arise in the cluster calculations, the latter are extreme groups within a cluster. Figure 8 overviews the approach of this study. To generalize within boundaries, a reference study is replicated and comparisons are given to the segmentation literature at different stages ("benchmarking").

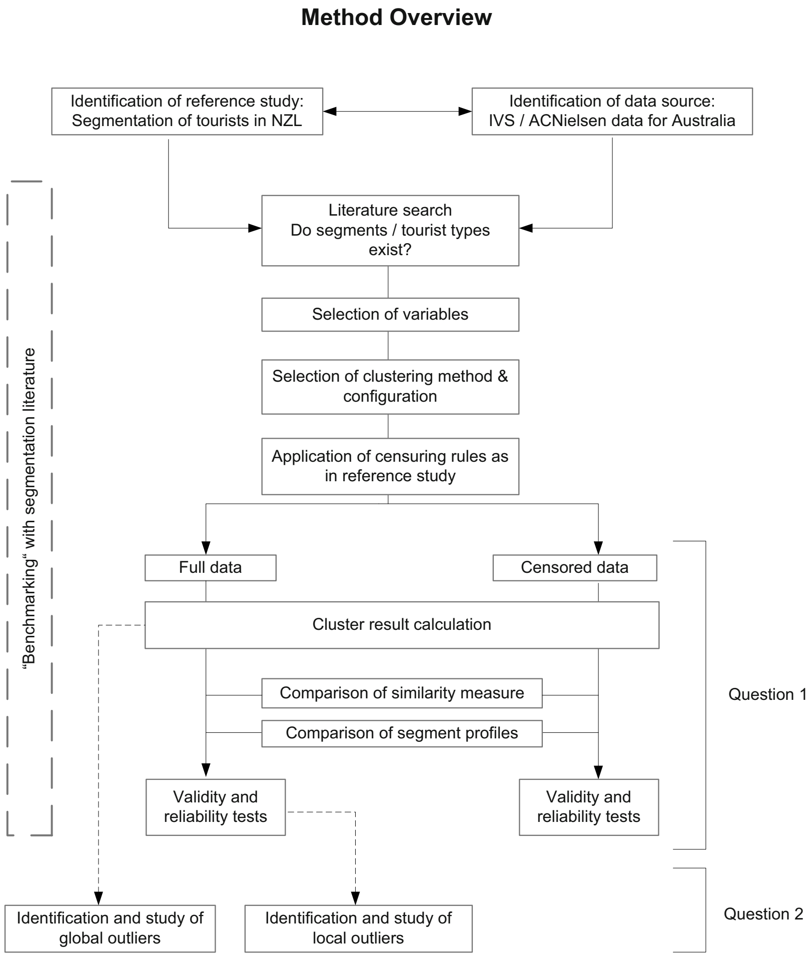


Figure 8: Method overview

2.1. Sample description and data collection

Data for this study were provided by the Australian Bureau of Tourism Research. AC Nielsen was hired to survey a representative sample of the Australian visitors by interviewing about 20,000 visitors each year. The "International Visitor Survey" (IVS) is conducted according to a sampling plan of all commercial flights from and to Australia. This sampling plan is derived from the immigration cards every traveller has to complete when entering Australian territory. The IVS has been operating since the early 1970s and is conducted by computer assisted personal interviewing (CAPI) in the departure lounges of the eight major international airports: Sydney, Melbourne, Brisbane, Cairns, Perth, Adelaide, Darwin and the Gold Coast. Interviewers are well-trained and native speakers in one of the six languages for the interviews. Questionnaires were translated and back-translated by native speakers several times to assure the same meaning of the interview questions across different languages. This study uses data from the 3rd and 4th quarter of 2000. The survey methodology presented here has been adapted in the years that followed and may not fit completely with the approach used for today's IVS.

Detailed IVS information includes the behaviour tourists exhibit in Australia, such as accommodation, transport and activity decisions, their expenditure on a broad array of categories, pre-travel planning as well as demographic details. While the full questionnaire comprises 89 questions, *Figure 9* overviews the most relevant areas of the questionnaire for this study.

Figure 9: IVS Questionnaire content, systematized

The questionnaire is aimed at recording behavioural data at the end of a tourist's stay in Australia. To achieve a high degree of reliability and comparability across cases, proof (e.g.

tickets or bills) as well as cross-check questions are used at various stages in the recording process. Cards stating the available options and maps are shown to the interviewee as memory hooks. A grid with different categories is used to record expenditure.

IVS data are used because of the similar structure and survey method compared to the sparring study on New Zealand visitors. However, some differences apply:

Becken and Gnoth (2004) pre-segmented their dataset: Long term tourists (staying 180+ days), gateway only tourists (those who never leave the airport), cruise ship tourists and touring tourists (staying at least at two different locations) were identified first and focus was then given to touring tourists only. No reason is given for these sample limitations; one assumption investigated later is that homogeneity is increased. In the second part of their study, the authors identified and tested decision rules to identify individual segments ("tourist types") in the new dataset. This part is not replicated because this study only aims at demonstrating one procedure at a time.

Travellers declaring that they were visiting Australia for holiday reasons or to visit friends and relatives enter into this analysis. A total of 5173 cases are included. For the clustering procedure, seven variables describing the accommodation and eight variables describing the transport decisions are used.

Accommodation variables: Percentage of nights spent in…	Transport variables used: Percentage of transport using…
Hotel or resort	Car or 4WD
Backpacker hostel	Taxi, chauffeur, hotel shuttle
Bed and breakfast (incl. farm)	Domestic air
Home (incl. VFR)	Long distance train
Camper	Long distance coach
House	Ship, boat or ferry
Motel	Rental car
	Van

Table 31: Variables selected for clustering procedure

In comparison to the reference study, the backpacker bus category is extended over other long distance bus companies, the cook strait ferry, motorcycle and cycle variables are left out because of non-existence. Bed and breakfast accommodations are merged with farms because of the similarity of tourist profiles staying in a "near to family" accommodation. Rented houses are added as those are distinct from homes where visitors stay for free.

2.2. Analysis

In the reference publication, Becken and Gnoth (2003) conducted a segmentation study to identify travel styles by means of looking at activity, transport and accommodation consumption patterns. In a similar way, the present study also identifies literature on travel styles in Australia first (e.g. Riege and Perry 2000, Horneman, Carter et al. 2002). Then, the distribution of potential cluster variables is inspected and – similar to the literature study – leading to the conclusion that clusters exist in the data space. Such tourist types in Australia include backpackers (Loker - Murphy and Pearce 1995, Pearce 2007), package tourists (Pan and Laws 2002), senior travellers (Horneman, Carter et al. 2002), travellers visiting friends and relatives (VFR) (Moscardo, Pearce et al. 2000) and educational travellers (Ritchie, Carr and Cooper 2003). For validation purposes, these types are used in the results section.

Pre-calculation stage

Before starting with the main analysis, many published studies have been subjected to other pre-processing analyses: 26.7% used factor analysis, 9.2% standardisation, 4% conjoint analysis and 7.2% other techniques. Additionally, the authors of 7% of the publications checked the data structure, identified literature suggesting the theoretical existence of a cluster structure in the data or pre-test and manually or visually inspected data. Only 44.6% of the published studies do not use any pre-processing at all. The remaining share is split into the use of multiple pre-processing techniques (Baumann 2000). Becken and Gnoth's study locates publications which indicate the existence of groups of tourists with similar behaviour. In terms of pre-processing, activity variables (e.g. beach visits, shopping, outback visit) were subjected to factor analysis before they were added to transport and accommodation variables for clustering. Unfortunately no detailed information concerning the factor analysis is given.

From a statistical point of view, data pre - processing such as standardizing and factor analysis have been criticized because… "part of the structure (e.g. dependence between variables) that should be mirrored by conducting cluster analysis is eliminated" (Arabie and Hubert 1994). Pre-processing the variables using factor analysis was considered for this study as well to stay as close as possible to the reference study and to decrease the number of variables. This would make sense if it renders the analysis more structured, particularly as thirty different types of activities were recorded in individual variables. Nevertheless, several arguments weigh heavier and in favour of an exclusion of those variables:

Summarizing activity variables from the content proved to be difficult and ambiguous.

Pre-processing through factor analysis is considered problematic from a statistical side and from an epistemological view as discussed before. The main argument is that using factor analysis or PCA (= Principal component analysis) to 'reduce' data prior to running a cluster analysis discards relevant 'distance' information. 'Tandem' clustering (i.e. the combined use of cluster and factor analysis) is thus often considered "outmoded and statistically insupportable" (Arabie and Hubert 1994, Ketchen and Shook 1996).

The decision to exclude activity variables from clustering is also supported to some extent by Becken and Gnoth (2004) who stated that the factor scores derived from the number of visits at specific activities played only a minor role, while discriminating variables could be found in transport and accommodation variables.

Finally, the exploratory factor analysis pursued for test purposes showed a "weak" solution (see Exhibit 8).

Standardization is also often discussed for variable pre-processing and the literature holds different views about whether variables should be standardized or not. On the one side, standardization can be important to bring variables with substantial differences in variance "on the same level". Otherwise variables with larger scales or variances will excessively influence the analysis. On the other side, standardization can destroy an existing structure in the dataset. If clusters are believed to exist in the transformed space, variables should be transformed.

Becken and Gnoth's study does not make any statement which would allow inferences about standardization; the present study therefore also abstains from standardization.

Principal component analysis on activity variables

A principal component analysis with Varimax rotation and Kaiser normalization was used for computation, after inspecting the correlation matrix for singularity and multicollinearity. Both checks performed positively (i.e. no bivariate correlations higher than +/- .9 and the determinant higher than .00001) (Field 2000). The KMO measure of sampling adequacy (MSA criterion) revealed a value of .87, which exceeds the recommended minimum of .80 for PCA (e.g. Backhaus, Erichson, Plinke, and Weiber 2003). The Bartlett test for sphericity yielded a highly significant result ($p < .001$). Factors were derived for Eigenvalues >1. Seven factors were extracted explaining 47% of variance, which is rather low as 53% of the information gets lost through this procedure. For the description of the factors, factor loadings larger than .34 were used and independently assigned and described by a second researcher familiar with the research domain. Factor calculations showed stability and were congruent with factor calculations in sub - samples and after the elimination of variable with weak loadings.

The Cronbach Alpha investigation for reliability (Peterson 1994) resulted in only two factors above a threshold of .70. Therefore, activity factors were excluded from the following cluster analysis procedure and the raw data were used for descriptive purposes.

Factor / Description	Highly loading variables
(1) Individualistic, not country specific leisure activities in urban surroundings (Alpha .73)	Visit botanical gardens, go shopping (pleasure), visit history/heritage buildings, go to markets, go to beach, visit zoos, aquariums, w. parks, visit nat. parks, visit museum/art gallery
(2) Explore native Australia (aboriginal) and nature (Alpha .71)	Outback activities, experience aboriginal art and craft, other outdoor activities, visit aboriginal site/community, tourist trains, guided tours and excursions, whale watching
(5) Consumption of cultural and sport events (Alpha .48)	Attend festivals/fairs/cultural events, attend sport events, attend theatre/concert/performing arts
(4) Experience manufacturing of traditional Australian products (Alpha .51)	Visit wineries, visit industrial attractions, visit art/craft workshop, visit farms, whale watching
(3) Commercially guided leisure excursions (Alpha .44)	Visit theme parks, guided tours and excursions, wildlife parks/zoos/aquariums, visit friends (negative), go to pubs& discos (negative)
(6) Non main stream activities (Alpha .22)	Fishing, other sports (not explicitly mentioned)
(7) Elite Activities (Alpha .11)	Visit casinos, golfing

Table 32: Factor Analysis Solution for "Activities"

Exhibit 8: Factor Analysis for activity variables

Method selection

Most segmentation studies use clustering methods (93.6%) for segment calculations. Hierarchical and partitioning methods have almost equal shares, while fuzzy clustering, optimization and other methods only make up the remaining 10%. For hierarchical procedures, agglomerative methods are preferred based on Ward's minimum variance method (56.6%) or complete linkage (9.6%). Other agglomeration methods only account for 7.2% each, or less, of the publications in this section. In the area of partitioning procedures, k-means represent 75.6% of the studies, while 18.9% do not specify the algorithm and 5.5% use very specific ones such as "Reloc", "CL" clustering, self organizing feature maps (e.g. Kohonen 2001) or latent class modelling (e.g. Goodman 1974, Lazarsfeld and Henry 1968). Euclidean Distance is used by 95.7% of the studies stating the proximity measure (Baumann 2000).

In the reference study, Becken and Gnoth (2004) considered a single hierarchical cluster analysis based on an average linkage algorithm not feasible. Therefore, the dataset was randomly split into four parts and the analysis was conducted separately for each of these parts. In the section that followed, the four parts were recombined, but no details were given for this step. Interestingly, average linkage methods are sometimes considered inferior to Ward and β-flexible methods to unveil the correct cluster structure (see e.g. Milligan and Hirtle 2003 for an overview on simulation studies). One reason may be the above average sensitivity of e.g. Ward methods towards outlying data. The present study uses the same algorithm as the original. Because calculation power permits, splitting the dataset is avoided and a single calculation is favoured to avoid merging errors. As in the reference study, a correlation similarity measure is used. Becken and Gnoth (2004) claimed that this shape measure is more appropriate to discover patterns in the data. Finally, as discussed previously, the match of the epistemological assumptions with the selected method needs to be investigated. Exhibit 9 details the fit of hierarchical cluster analysis with the paradigm of critical realism.

Epistemological foundations of cluster analysis

As discussed in Chapter 3, not all statistical procedures match the assumptions of a critical realist approach. Despite the existence of a wide array of techniques to derive groups, only those based on cases may be considered "because the probabilities are the ontological characteristics of the situation and not an epistemological tool to stand in for a lack of knowledge" (Williams and Dyer 2004, p78). For instance, many researchers combine factor analysis with clustering to reduce the amount of variables. In a realist view, this approach is questionable because "factor analysis can be used to demonstrate multiple contingency amongst highly inter-correlated variables, but such techniques are epistemological techniques and can still provide only inference to the best explanation" (Williams and Dyer 2004, p78). Cluster analysis itself is essentially taxonomic and was originally conceived to estimate the degree of similarity among organisms and eventually group them (Aldenderfer and Blashfield 1984). Unlike variable-based analysis, cluster analysis departs from the case and looks for similarities between cases.

When travellers are interviewed, they look back on a history of prior events (i.e. accommodation and transport decisions). Some have had similar antecedents, some not. These events have influenced their behaviour. At a number of measurable points, decisions have been made which either similar to others or not. These points are now used for comparison and also represented in the clustering structure: Each case is first put in its own cluster and at each level of the analysis, clusters are combined in a traceable way which can also be represented in e.g. a dendrogram. As the number of clusters decreases, groups get increasingly dissimilar, reflected by the error sum of squares or other measures. Plotting those, a sharp increase also signals a sharp decrease in homogeneity in the derived solution and signals that much of the accuracy has been lost by reducing the number of groups (Williams and Dyer 2004). The approach described here is valid for hierarchical clustering, however, the procedure used in this work – a combination of hierarchical and k-means clustering – can be considered as fitting the realist approach as well:

* k-means procedure (SPSS quickcluster algorithm) is used to classify the cases to existing cluster means derived from hierarchical calculations

* cluster solutions at the end are compared in terms of stability (overlap of two solutions derived from a split sample calculation)

Exhibit 9: Epistemological foundation of cluster analysis

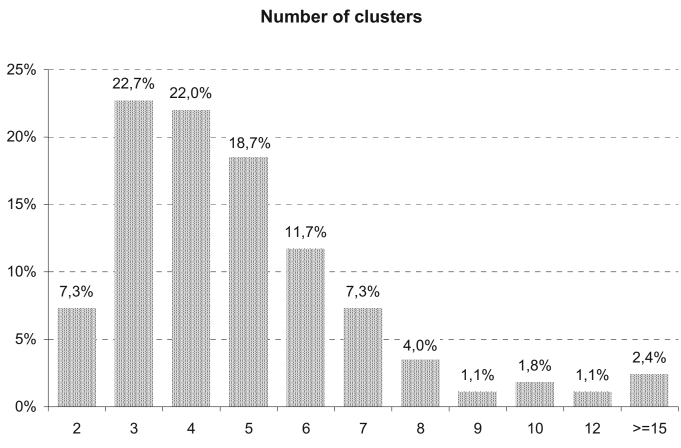
Variable selection and weighting

Weighting variables is often considered worth discussing when using cluster analysis, particularly if some clustering variables are suspected of contributing particular noise (see e.g. Milligan and Hirtle 2003 for a discussion). As the original study did not use any weighting and no meaningful weighting procedure could be derived for the approach taken here, this study abstains from doing so in order to keep the results comparable.

Finally, accommodation and travel choice variables were created summarizing similar types of usage (e.g. hotel and resort stays or taxi and shuttle service are combined) and normalised as presented in the reference study: Each traveller was attributed a percentage of nights spent in a particular type of accommodation and a percentage of travel decision taken in favour of other means of transport.

Selection of desired number of clusters

To determine the number of clusters, about 50% of the publications declare using a specific method, 23.6% combine method and subjective choice, 19.7% do not give a reason for their choice at all and 6.7% use subjective means only (e.g. the interpretability of results). The methods applied to determine the "right" number of clusters vary to a large extent: Graphical representations, hierarchical methods, cubic clustering criterion, variances, F values and other statistical tests, as well as theory-based clusters are the most prominent ones (Baumann 2000). Most of the studies derived between three and six clusters as shown in Figure 10.



Source: Baumann 2000

Figure 10: Number of clusters derived in segmentation studies

The reference paper takes an exploratory approach towards determining the number of clusters and derives the number of clusters and the comparison of cluster means in a seemingly intuitive way. The number is derived by comparing possible solutions from a dendrogram. Unfortunately, this dendrogram is not shown. To complement this approach, the study presented here uses stopping rules: Milligan and Cooper (1985) have extensively simulated different stopping rules and proposed the use of two or three of the better performing ones (Milligan and Hirtle 2003). In the following study, the two "best" performing rules – the Pseudo - F (Calinski and Harabasz 1974) and Pseudo - T^2 (Duda and Hart 1973) – are used in the way they are implemented in the software package Stata SE 10. Exhibit 10 explains the background and the handling of those rules.

Stopping rules for cluster analysis
The Pseudo - T^2 statistic (PST²) is derived from Hotelling's T^2 test. High Pseudo - T^2 values indicate high differences in the mean vectors of the two clusters just combined or – put differently, the Pseudo - T^2 is the ratio of the increased within-category variance after joining two clusters to the variance within each of the two clusters. If the value Pseudo - T^2 is small, the researcher assumes that clusters can safely be combined (Duda and Hart 1973, SAS Institute Inc.). The Pseudo - F statistic (PSF) (Calinski and Harabasz 1974) measures the separation among the clusters at the current level in the hierarchy. Large values indicate that the mean vectors of all clusters are different, while F is not distributed as a random variable. High values, particularly those preceded and followed by lower values signal potential candidates for cluster solutions. In other words, the Pseudo - F is the ratio of the between-cluster to within-cluster variance (SAS Institute Inc.).

Exhibit 10: Stopping rules for hierarchical cluster analysis

Validation and interpretation

Fifty percent of the segmentation studies did not state any details concerning the validity; others calculated some kind of significance tests (26.9%) or pursued a discriminance analysis (12.2%). Similar results exist for reliability testing: 66.7% did no reliability testing at all; 14.6% pursued a split-half approach; while the remaining options were used by less than 5% of the publications (Baumann 2000).

In the reference study, Becken and Gnoth used ANOVA and Chi² statistics to test the results. Because of the nature of the cluster analysis, which maximizes mean difference on the clustered variables between the clusters, this internal consistency measure has been frequently criticized. This is particularly obvious in the case of a minimum variance ("Ward") method, which uses a procedure similar to an ANOVA to derive the cluster solutions (Milligan and Hirtle 2003). The present study abstains from this practice and additionally validates the results with descriptives from cluster variables, non-cluster variables and externally by means of other studies.

To test the reliability of the solutions found, this study uses a split-half validation (McIntyre and Blashfield 1980; Morey, Blashfield and Skinner 1983). The logic of this approach suggests that if an underlying clustering exists in the dataset, then one should be able to replicate these results in a second sample from the same source and set of variables. Therefore, the present sample is randomly split into halves. The first sample is clustered in the planned manner, then the centroids of this first halve are used to determine the distances to the cases of the second halve. Those are then classified in the respective groups using a k-means procedure with the same distance measure as applied in the hierarchical clustering. Finally, the planned clustering method is also applied to the second half of the sample. The agreement of the k-means attributed cases and the clustered cases reveal a measure of reliability for the clustering undertaken. Milligan and Hirtle (2003) propose the use of a Kappa statistic for this purpose.

3. Results

In this section, the results of a calculation using the entire sample based on 5,173 cases ("full data") and the calculations on the censored data ("censored data") based on 2,406 cases are presented. Furthermore, outliers and their additional gain in information are investigated. The software suite Stata SE 10 served as the platform for the calculations. Sub-section 1 presents the results for the research question

Does a priori elimination of outliers and extreme cases through e.g. sample censoring or pre-segmentation yield different results compared to calculation with the entire data?

Sub-section 2 presents the results for the research question

Does studying outliers and extreme cases result in additional theoretical insights?

3.1. Impact of censoring on the results of a study

Table 33 displays the effect of censoring on the sample size within the present study. The reference study pre-segmented the data by eliminating long-term travellers (> 180days), cruise ship tourists (those who spend their nights on a cruise ship) and non-touring tourists (those who stay in less than 2 different places). While the reference study reported a total reduction of the original sample of 33%, the application of the same criteria resulted in a reduction of 53%. Most of the reduction can be attributed to the criterion “Stayed in at least two different locations”, as shown below.

All Cases	Stayed more than 180d	Cruise ship tourists	Stayed in less than 2 locations	All rules applied
5173	- 100	- 188	- 2430	2406
All reductions calculated on initial dataset. Gateway tourists were not covered by the dataset in use				

Table 33: Reduction of original dataset through censoring rules

The first decision a researcher takes after entering the variables in a software interface is the number of groups to derive. While the reference study used an exploratory approach (i.e. skimming over possible solutions around five groups), this study investigates the dendrogram and additionally applies stopping rules to allow a more detailed comparison. Inspecting the dendrogram, one outlying group can be identified and is investigated in section 3.2. The exclusion of this small group (n = 5) has been simulated by and resulted in small within-profile changes, but medium scale changes in terms of the profile share of the whole dataset. Appendix 2 exhibits those changes more in detail. The extreme group is kept in its original form in the dataset for the subsequent analyses. Figure 11 and Figure 12 present the Pseudo - F and Pseudo - T^2 scores as proposed by Milligan and Hirtle (2003). For the interpretation of these graphs, one reads the plots from the right to the left side. A high Pseudo - T^2 statistic indicates that the mean vectors of the clusters that are combined in the next clustering steps are highly different and thus should not be combined. Therefore, one seeks peaks of this statistic and takes one or two steps back to where the score is small and thus the combination of clusters is safe. At the same time, at the given cluster number, the separation of the clusters should be as high as possible. This is tested using the Pseudo - F score: A local peak of this score indicates a high level of separation between the clusters, which is desirable.

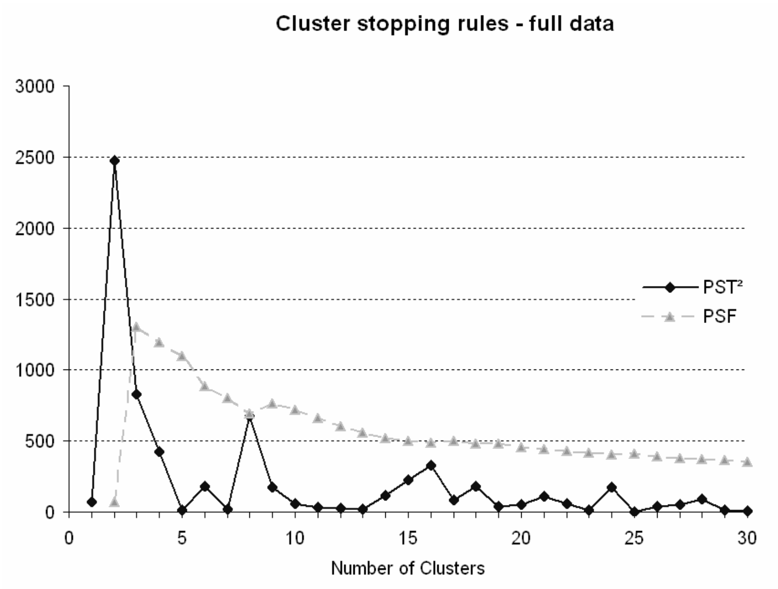


Figure 11: Cluster "stopping rules" for full data

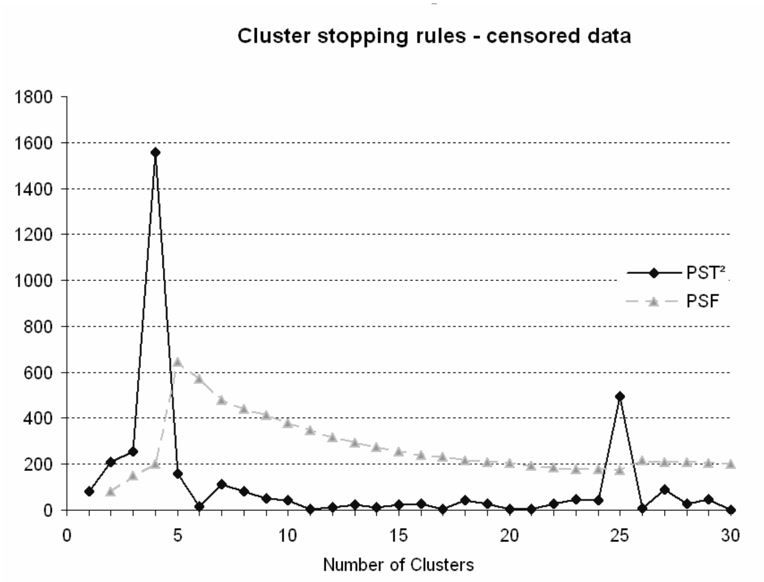


Figure 12: Cluster stopping rules for censored data

Seeking possible cluster solutions in the full sample data, one notes that the full sample solution would suggest the calculation of three, four or five groups. As a second best solution, nine or ten clusters can also be considered. Looking at the censored sample, one would highly favour a five or six cluster solution or – if a much higher number of clusters is acceptable – a 26 cluster solution.

Using the stopping rules as the decision criterion for the number of clusters to derive, a clearer decision can be made for the censored sample than for the full sample. This applies particularly for up to 10 clusters, which is also the relevant area where more than 95% of the segmentation studies (Baumann 2000) find their solution.

In the present study, a five cluster solution is derived for both datasets in order to facilitate comparison. This is consistent to the general tendency of segmentation studies in top tier journals, which retain an average of 5.2 groups (Franke, Reisinger and Hoppe 2007). Of course, additional information could be gained by inspecting more groups. Potential ideas solutions for this approach may be found in Figure 11. However, this chapter follows the path presented by the reference study. As the results of a cluster analysis depend highly on the algorithm chosen as well as data properties, the researcher is advised to take care when extrapolating the results to other analysis techniques.

The Pearson correlational similarity measure expresses the similarity between different groups in a cluster solution and thus the heterogeneity at a given number of clusters. This similarity measure was used for both the reference study and the present study. Highly positive values (i.e. close or equal to 1) signal the similarity that would be attained if every case were to be treated as an individual cluster. If all cases are treated as one single group, the correlational measure approximates 0. The measure itself may range from -1 to 1. Figure 13 displays the development of the similarity measure for the two different datasets. As the cluster procedure undertaken in the study does not directly aim at minimizing within-cluster variance (i.e. absolute differences are not relevant), but to identify patterns (a "shape measure" as described in the reference study (Klecka 1980)), only between-cluster similarity is evaluated. Calculations reducing within-cluster heterogeneity used, for example, Euclidean Distance Measures combined with a Ward Method (see e.g. Franke and Von Hippel 2003 for comparison).

Figure 13 shows that up to ten cluster solutions, data censoring results in a substantial increase in the similarity measure. This means that sample censoring reduces heterogeneity; cases and, consequently, clusters built upon them become more similar. Interestingly, both lines converge between 10 and 15 clusters and then slightly separate again. Simulations with a reduced set of variables showed a very similar behaviour up to 10 clusters.

For the present case, one can state that the statistical similarity of cluster solutions is increased by censoring the dataset, particularly in the area up to 10 clusters. As the present study is similar to many other segmentation studies, one may assume a similar effect there. However, replicating this study on other datasets would be necessary to support the assumption

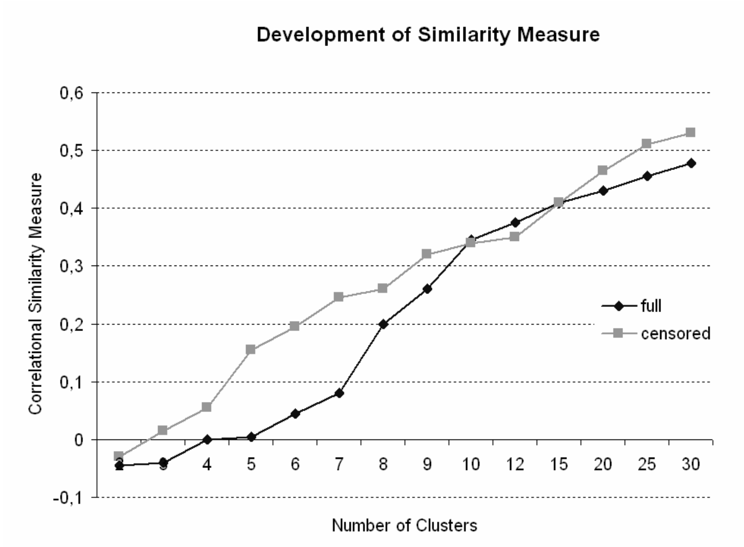


Figure 13: Similarity measure compared for the two solutions

The next four tables (Table 34 - 37) display the consumption profiles in a similar way to the reference article. Detailed results on the clustering can be found in Appendix 2. For the cluster variables, the most prominent ones have been selected using a threshold of 15% to increase readability. Tourist type labels have been used for similar profiles.

Tourist type	N / %	Travel choice variables used in cluster analysis	Descriptive variables for each tourist type
1) VFR	2189 / 42.3%	60% car 22% airplane 91% family or friend's home	21% NZL, 14% UK, 7% JAP 7% USA 72% return visitors, 99% no group tourists, 69% unaccompanied traveller Age: 40 yrs, 61% female Total expenditure: 1792 $, expenditure per day: $129 Stay: 24d
2) Backpacker	345 / 6.7%	32% coach 16% airplane 79% backpacker hostel	26% UK, 11% JAP, 9 % GER 76% first time visitors, 97% no group tourists, 64% unaccompanied traveller Age: 27 yrs, 53% male Total expenditure: $3707, expenditure per day: $116 Stay: 52d
3) Comfort tourist	2206 / 42.6%	36% airplane 25% charter bus 19% taxi or shuttle 90% hotel or resort	37% JAP, 9% USA, 8% NZL 63% first time visitors, 37% group tourists, 47% travelling couple Age: 37 yrs, 51% female Total expenditure: $2547, expenditure per day: $350 Stay: 9 d
4) Housestay & charterbus tourist	325 / 6.3%	32% charter bus 17% airplane 17% rental car 85% house	37% NZL, 12% SIG, 11% UK, 8% JAP, 36% first time visitors, 95% no group tourists, 38% travelling couple , 35% unaccompanied Age: 42 yrs, 57% female Total expenditure: $5185, expenditure per day: $198 Stay: 63 d
5) Camper	108 / 2.1%	43% van 21% rental car 70% camper	34% GER, 20% UK, 8%NED, 7% SUI 58% first time visitors, 96% no group tourists, 48% travelling couple Age: 37 yrs, 54% male Total expenditure: $5428, expenditure per day: $164 Stay: 39 d

Table 34: Tourist type profiles – full sample

For the calculations using the full sample, a four- and a six-cluster solution were also calculated. If four clusters were derived, the algorithm merges cluster three and four shown above. If six clusters are derived, five people are split apart from cluster three. They exhibit very high usage rates of ships as means of transport and correspond to the outlier profiles described in the next section in Table 39.

Table 35 is aimed at enriching and validating the segment profiles. The average number of visits in the last ten years not only reflects behavioural loyalty and habits, but also experience with the destination and its tourist offerings. The number of different information sources consulted before travelling is a proxy for the lack of prior knowledge, their nature reflects availability and trust in the respective source. The activities undertaken complement the consumption profile drawn previously.

Tourist type	Visits in last 10 yrs; Number of info sources; Mostly used sources	Mean number (SD) of activities realised; Popular activities (threshold 50%)
1) VFR	3.7 (5.9); 7 (1.2); Family/friend (58%)	6.9 (3.8); VFR (92%), Shopping (89%), Beach (61%), Pubs&Clubs (54%), Markets (53%), Botanical gardens (50%)
2) Backpacker	1.8 (2.7); 2.2 (1.8); Book /guide (71%), Travel agent (46%)	10.6 (5.8); Beach (80%) Nat. park (75%), Pubs&Clubs (70%), Botanical Gardens (63%), Wildlife parks (62%), Markets (61%), Guided tours (54%), History& heritage buildings (53%)
3) Comfort tourist	3.8 (7.2); 1.9 (1.6); Travel agent (58%), Book/guide (57%)	7.6 (3.9); Shopping (90%), Beach (70%) Wildlife park (62%) nat. parks (59%), Guided tours (56%), Botanical gardens (51%) Markets (50%)
4) House stay & charterbus tourist	1.4 (1.0); 1.7 (1.6); Travel agent (60%), Book/guide (49%)	8.2 (5.2); Shopping (90%), Beach (84%), Markets (59%), VFR (53%), Nat. Parks (50%)
5) Camper	1.9 (1.8); 2.6 (2.1); Guide/Book (73%), Agent (59%), Internet (52%)	10.2 (5.1) Nat. Parks (87%), Shopping (80%), Beach (77%), Wildlife Parks (69%), Botanical. gardens (62%), Markets (58%), Outback (51%), Pubs&Clubs (51%)

Table 35: Additional characteristics of tourist types – full data

Tourists visiting friends and relatives do not require much additional information search as they stay with a trusted family member or friend. The opposite is true particularly for backpackers, who often plan their route themselves, and campers. Guide Books, such as those from the Australian Company "Lonely Planet" are very popular with tourists. Comfort and charter bus tourists can largely rely on booking information provided by their travel agent. Camper tourists may find it most difficult to get to the best places and thus refer to a wide variety of different sources.

In terms of activities pursued, all profiles include popular tourist activities such as beach, shopping and market visits. However, tourist type-specific differences exist. The VFR tourist generally partakes in a low number of activities. Backpackers also appreciate nightlife and guided tours, which are often small group tours operated by local tour guide offices to remote sites such as the “Great Ocean Road” in Victoria or “Daintree Rainforests” in Queensland. Comfort tourists also rely on guided tours and enjoy Australia’s nature, something that is included in many tour packages. However, a guided tour might have a completely different character including 30+ bus tourists compared to a tour involving 10 or less young backpackers. House stay and charter bus tourists exhibit similar patterns, but also frequently visit friends and relatives. Campers are particularly attracted by nature-based activities such as visiting natural parks, wildlife parks and the outback. Table 36 displays the censored version of the analysis.

Tourist type	N / %	Travel choice variables used in cluster analysis	Descriptive variables for each tourist type
1) VFR	788 / 32.7%	49% car 31% airplane 77% family or friend's home	21% UK, 18% NZL, 8% USA, 64% return visitors, 98% no group tourists, 64% unaccompanied traveller Age: 41 yrs, 59% female Total expenditure: $2420, expenditure per day: $111 Stay: 30 d
2) Comfort tourist	1189 / 49.4%	53% airplane 22% charter bus 88% hotel or resort	40% JAP, 10% USA 73%first time visitors, 45% group tourists, 55% travelling couple Age: 37 yrs, 54% female Total expenditure: $2880, expenditure per day: $327 Stay: 11 d
3) Self drive tourist	202 / 8.4%	83% rental car 26% motel 20% family or friend's home 15% hotel or resort 15% house	23% NZL, 16% UK, 15% GER, 58% return visitors, 98% no group tourists, 48% travelling couple Age: 43 yrs, 53% male Total expenditure: $4403, expenditure per day: $260 Stay: 21d
4) Backpacker	164 / 6.8%	47% coach 20% airplane 79% backpacker hostel	24% UK, 9% NED, 9% GER 77% first time visitors, 98% no group tourists, 69% unaccompanied traveller Age: 27 yrs, 53% female Total expenditure: $3508, expenditure per day: $108 Stay: 40d
5) Camper	63 / 2.6%	63% van 69% camping car	40% GER, 19% UK, 13% SUI, 11% NED 62% first time visitors, 97% no group tourists, 49% travelling couple Age: 36 yrs, 51% male Total expenditure: $4899, expenditure per day: $143 Stay: 35d

Table 36: Tourist type profiles – censored sample

In case a six cluster solution is calculated for the censored sample, cluster two is split and 52 people form a new group, mainly using (rented) houses for accommodation and airplanes for transportation. Table 37 complements the profiles from previous tables.

Tourist type	Visits in last 10 yrs; Mean # (SD) info sources; mostly used ones	Mean number (SD) of activities realised; Mostly realised activities (threshold: 50%)
1) VFR	3.7 (6.0); 1.1 (1.5); Family/ Friend (58%)	8.5 (4.4); VFR (91%), Shopping (88%), Beach (70%), National parks (59%), Markets (58%), Pubs&Clubs (58%), Botanical Gardens (55%)
2) Comfort tourist	1.8 (2.7); 2.1 (1.5); Travel book/guide (58%), Travel agents (56%)	8.5 (3.8); Shopping (92%), Beach (75%), Park, Zoo, Aquarium (74%), National parks (73%), Guided tours (70%), Botanical Gardens (61%), Markets (53%)
3) Self drive tourist	3.8 (7.8); 1.9 (1.8); Travel book or guide (59%)	8.7 (3.9); Shopping (84%), Beach (76%), National parks (75%), Botanical Gardens (58%), Markets (58%), Wildlife parks (55%), VFR (55%), Pubs&Clubs (50%)
4) Backpacker	1.4 (1.0); 2.1 (1.7); Travel book / guide (69%)	11.4 (5.1); Beach (85%), Shopping (82%), National parks (81%), Pubs&Clubs (74%), Markets (70%), Botanical Gardens (68%), Wildlife parks(68%), Guided tours (63%), History/heritage building (56%), Museum/Art(56%) Other outdoor activities (52%), VFR (50%)
5) Camper	1.9 (1.8); 2.8 (2.1); Travel book / guide (75%), Travel agent (56%), Internet (54%)	10.6 (4.7); Nat. Parks (92%), Shopping (82%), Beach (78%), Wildlife parks (78%), Outback (67%), Botanical gardens (63%), Markets (60%), Pubs&Clubs (51%)

Table 37: Additional characteristics of tourist types – censored data

Similar to the full sample table, VFR tourists use little information sources and refer to friends or relatives, while the other groups, particularly campers, use a much higher variety of sources. Travel books and guides are the most popular information sources, followed by travel agents. In terms of loyalty, VFR and self-drive tourists visited Australia much more often than other tourist types. VFR tourists are the least active in this calculation and mainly differ to other profiles by enjoying nightlife and visiting friends. Comfort tourists experience more organized types of activities such as guided tours or parks, zoos and aquariums. Self-drive tourists mostly differ by their above average visits to national parks and also visiting friends and relatives to some extent. Finally, backpackers are likely to enjoy the nightlife, but also to use guided tours.

For the purpose of reliability analysis, the coefficient of agreement, “Kappa”, is calculated through cross validation, as proposed by Milligan and Hirtle (2003). The results are displayed in Table 38.

Full Sample	Censored Sample
Agreement: 83.09%; Kappa: .73; SE: .014 Z: 51.49; Prob. Z: .000	Agreement: 77.34%; Kappa: .64; SE: .019 Z: 33.73; Prob. Z: .000

Table 38: Results from replication analysis

The values reported above can be seen in the light of Breckeridge’s (1993) study who – in his simulation study of 960 error-free datasets – found an average of 74% of agreement for group-average methods.

The ANOVA calculated on all transport and accommodation variables, as well as length of stay and expenditure, showed significant group differences at the 1% level. Chi^2 tests on age groups, sex, country of residence, travel party size and previous visits to Australia were also calculated. The only exception is the variable “long distance train” which exhibits a between-group significance of p=.084, while age, travel party composition and expenditure per day yielded below 5%. One reason for this deviance in the censored sample may lie in the reduced number of cases in the censored data. In this case, the calculations of the significance tests are purely descriptive and aim to complete the replication (See critique in the method section of this study).

Overall, one would evaluate both cluster solutions as valid under current standards of analysis.

Comparing the result from the censored sample with the one from the full sample, one can see some striking differences:

The VFR type in the censored sample highly overlaps with its counterpart in the full sample. It, nevertheless, differs highly in absolute as well as relative size. While this tourist type makes up 42.9% (2217) of the full sample, it is reduced to 32.7% (788) in the censored sample. Generally, the existence of this type is supported by Moscardo (2000) or the reports from Tourism Australia (Batteye and Suridge 2003).

The comfort tourist type also exists in both samples and represents a similar share of the sample. Nevertheless, the ratios of Japanese travellers as well as the ratios of group travellers are significantly higher in the censored sample than in the full sample. To validate these findings, Pan (2002) identified travellers coming from Japan and other Asian markets who account for a high number of the package tours and many first-time visitors. In particular, Japanese travellers are attracted by the destination. They exhibit a low level of price sensitivity, which is reflected through above-average per day expenditures.

The self-drive tourist is a frequent phenomenon in the tourism literature as, for example, the "automobile tourist" in New Zealand (Becken and Gnoth 2004) or in Australia (Carson, Waller and Scott 2002). This tourist type, using rental cars and a wide variety of accommodation options, does not exist in the full sample. The opposite is true for the house and charter bus tourist type, a long-term tourist that only exists in the full sample.

The backpacker tourist type in the censored sample corresponds to segment two in the full sample. The main difference can be seen in the share of nationalities: The full sample covers a higher share of Japanese travellers, whereas the censored sample includes more Dutch people. Despite this small variation, backpacker tourists with similar profiles have been reported by, for example, Loker-Murphy (1995) as young budget travellers in Australia.

Finally, the camper tourist type also exists in both calculations. Nevertheless, its share in the censored sample is substantially higher. This type scores higher in total expenditure and length of stay in the full sample. The Camper tourist has been identified by Becken and Gnoth (2004) for New Zealand, but also in the publications of the Australian Tourism Research Association (e.g. Batteye and Suridge 2003). Interestingly, domestic Australian camper tourism is structurally different and includes, for example, mostly retired people (Prideaux and McClymont 2006).

Overall, one can summarize that four roughly overlapping profiles have been found in the full data, compared to the censored data and one profile with completely different characteristics emerged.

This “new” profile, the charter bus/house tourist appears to be somewhat special if only perceived via transport and accommodation patterns. Looking at the distribution of cases and their consumption systems, some differences to other groups arise. For instance, 36% of this group comes from the Pacific region (New Zealand, Pacific islands, etc.) and 37.8% are between 20 and 30 years old. Looking at Australia’s tourism industry, some charter bus companies use or even operate guest houses when taking particularly young travellers along popular tourist tracks for a week or two. Interestingly, 19.6% of the group also travel on a working holiday programme (mostly European and Japanese residents). If travelling with this programme, residents from many European countries as well as Canada, Japan, South - Korea or Hong Kong are allowed to work for an extended period of time in mostly labour-intensive industries, such as fruit picking, call centre operations or tourist services, in order to fund their holidays in Australia. Clarke (2004) describes this type from an ethnographical point of view, where he underlines, among other factors, the educational intentions of this agreement between foreign countries and Australia. The relatively high mean age in this group results from a substantial group of elderly New Zealanders who also used charter busses and houses for overnight stays and were therefore grouped there by the clustering procedure. At first sight this profile might appear error-laden. Looking at the within-case information, one is able to reveal and even validate the profiles from the literature.

In terms of validation with other published studies, the censored sample may largely represent the existing view of tourism in Australia.

This view is very similar to the results found in the reference study for New Zealand. For instance, the self-drive tourist has been identified as “automobile tourist” (Becken and Gnoth 2004) as well as in other Australia-focussed studies (see e.g. Carson, Waller et al. 2002), whereas the house/charter bus tourist does not appear explicitly and distinctly. Nevertheless, this may not only be attributed to the data per se, but also to the way this segmentation approach is pursued. The use of product-specific, non-observable characteristics (e.g.

psychographical or benefit measurement), for instance, has been shown to be very promising for deriving segments (Wedel, Kamakura et al. 2000).

As shown in this chapter, the reference study is typical for many of the studies pursued in the area of segmentation and tourism in particular. One notes that sampling, pre-segmentation and the exclusion of outliers may severely impact the results of the study. Each of these measures requires careful reflection and reasoning.

The relationships within the consumption systems framework are often investigated using correlation models. The relationships between the variables in the study were calculated similar to Woodside and Dubelaar's (2002) study to explore the existence of a consumption system. Even though the Pearson correlation coefficient signals a relationship, no indication for causality or a potential direction may be given from this exploration. The coefficients calculated for this study showed differences in correlations between the censored and the full sample for almost any relation. For example, the relationship between boat transport and accommodation changes r from .62 to .40. However, the great majority of differences between the correlations of the two samples are in the area of .1 or less. One can conclude that censoring has a small, but distinct, influence on the relationships between the variables in use.

3.2. Gaining insights by studying outliers

The procedure used in the previous section replicated Becken and Gnoth's (2004) study and demonstrated the effect of censoring on global outliers. This refers to outliers which arise from the cluster analysis procedure per se. This section draws on the same cluster analysis to investigate the outlying group and complements the study on outliers through an analysis on local (i.e. within-cluster) outliers.

Cluster analysis is used to reduce complexity in the dataset by forming a manageable number of groups. In the hierarchical clustering process, the decision to use a particular cluster solution is often guided by a dendrogram displaying the measure of similarity or distance between the grouped cases for solutions with different numbers of clusters. The dendrogram can also help to identify outliers. Gnanadesikan (1977) and Everitt, Landau and Leese (2001), for example, argue that individual cases or small groups, which are merged at a very late point in a hierarchical clustering process greatly influence the result of the analysis. The dendrogram in Figure 14 is derived from the calculation using the full sample in section 3.1. For clarity, 50 clusters have been plotted instead of the full dendrogram (5,173 branches). One group stands out as merging particularly late with the remaining data: Group 38 is characterized by (cruise-) ship users: On average, they stay 73% (SD .44) of their nights on a ship and use it for 92% (SD .14) of their transports decisions. Those were eliminated by one of the rules for sample censoring in the previous section. Further details are displayed in Table 39. Compared to what is known from cruise ship tourist profiles, this consumption profile appears to be in line with recent knowledge on cruising tourists (see e.g. Petrick 2004). One can therefore assume that the data are not based on errors.

Characteristics of outlier group	
Nationality: NZL (2), USA (1), HKG (1), GER (1) Age 47.2 (ME 52; SD 19.7); Sex: 80% female	60% return visitors; 80% no group tourist; 73% nights aboard; 92% of transport by boat
Travel party: 3 adult couple, 1 unaccompanied, 1 friend/family all holiday travellers	Av. Nights: 18.8 (Me 11; SD 20.0) Exp. / day: M $127 (ME $130 SD $98) Total exp. $2711 (ME $1728 SD $3556)

Table 39: Characteristics of outlier group in cluster analysis study

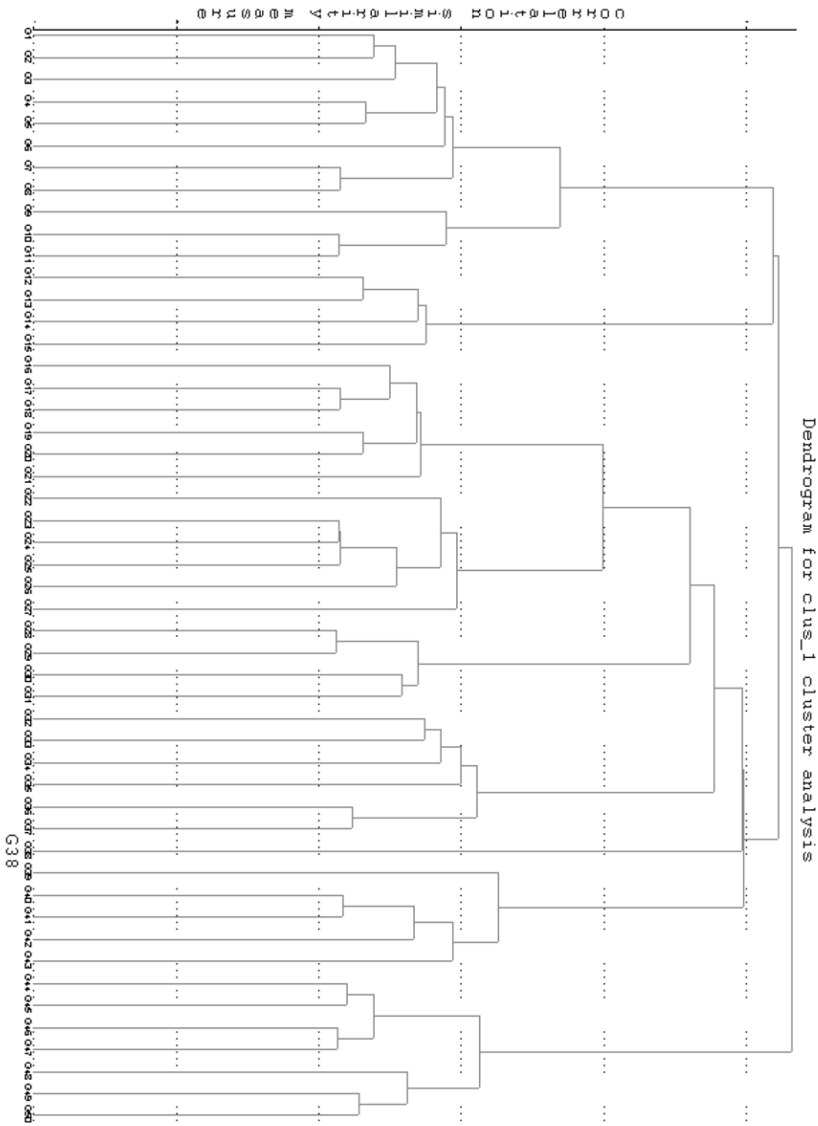


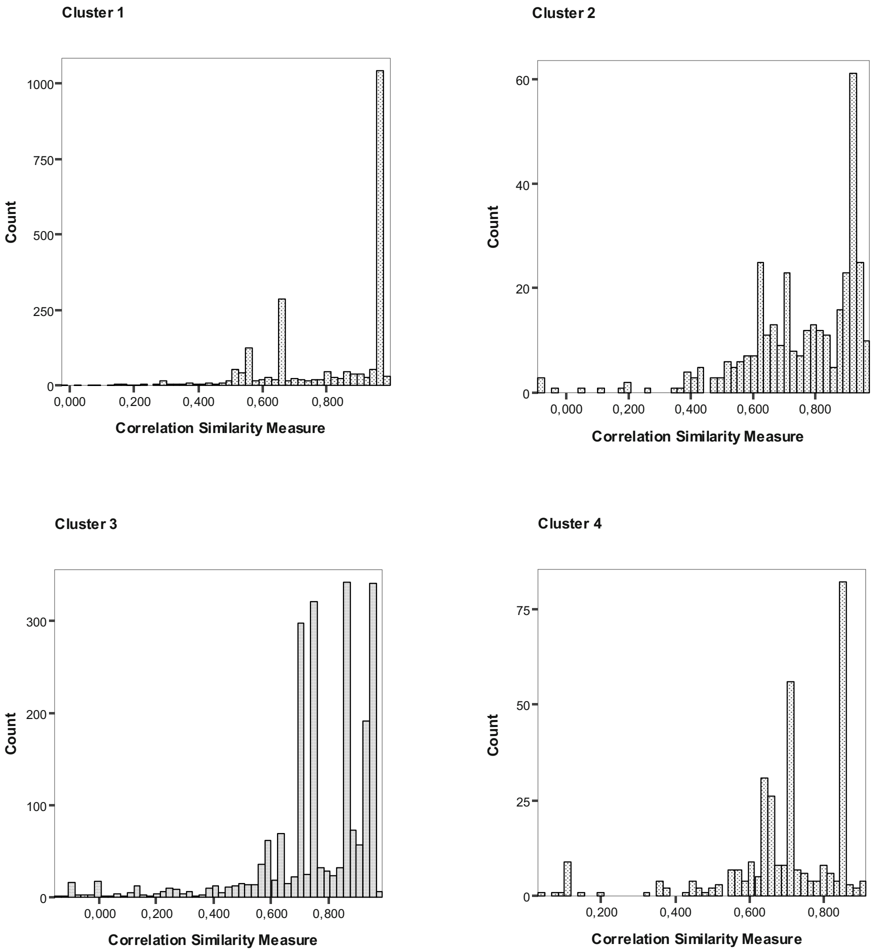
Figure 14: Dendrogram for full sample

Interestingly, cruise ship tourists have been studied by other researchers and evaluated as particularly relevant for countries such as Australia: Their number is undisputedly increasing and thus their economic impact (see e.g. Dwyer and Forsyth 1998). Because of the high variation in this group, it may be fruitful to distinguish between yacht tourists (for which no category in the questionnaire is provided) and cruise ship tourists. Furthermore, only a small proportion of all cruise ship tourists fall in to this extreme group.

Overall, one can summarize that the extreme group in this case is not due to error, as existing sources indicate the increasing importance of the cruise tourists has revealed.

Oftentimes, clustering can only reduce the variance in the dataset to some extent. Franke and Von Hippel (2003), for instance, report an average of 50% of variance remaining within the cluster after its calculation. This can be quite critical if one aims to address customer preferences with a small number of products (Franke and Von Hippel 2003). Local outliers may substantially contribute to this variance but, on the other hand, may also offer interesting insights in to, or examples for, combinations of customer needs or new needs that are arising. This subsection, therefore, proposes the investigation of local, within-cluster outliers.

In this investigation, outliers have been defined as those cases which substantially increase the within-cluster variation or which are highly dissimilar to the average cluster profile. In a first step, the distribution of the similarity measure for all clusters of the full sample is inspected. Those are calculated using the Stata 10 command, “cluster measures”, and are then printed in a histogram as shown in Figure 15.


Cluster 1
Count
Correlation Similarity Measure
Cluster 2
Count
Correlation Similarity Measure
Cluster 3
Count
Correlation Similarity Measure
Cluster 4
Count
Correlation Similarity Measure

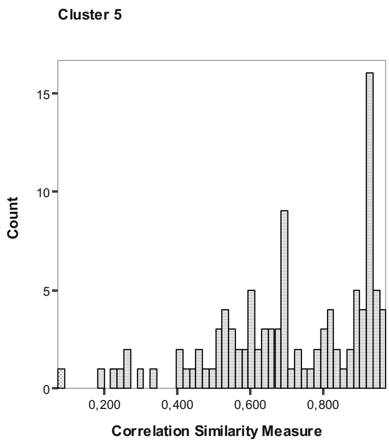


Figure 15: Plots of similarity measures

Not surprisingly, all figures display a concentration of cases towards the upper end (i.e. the cluster centre) of the scale. A more aggregated level of granularity (i.e. categories of 0.1 spread) may even underline this picture. Nevertheless, the distributions of the similarity measure for cluster two, four and five exhibit cases close to zero and thus dissimilar to the remaining group. Some researchers may interpret them as the "tail of a distribution". These cases sometimes represent groups apart such as in the plot of cluster four.

To analyse those cases, one needs to keep in mind that a similarity measure represents only an aggregated, univariate measure. Consequently, the extreme group may be scattered in the multivariate data space. As a solution to this problem, the subgroup could be clustered again. To illustrate the differences against the remaining data, however, other techniques may be more suitable, such as tree-based classification techniques which help to "predict" the source of extremeness of the distance measure by using a few variables. Exhibit 11, at the end of this section, discusses one of these techniques in more detail.

For the current calculations, an exhaustive CHAID technique was used to identify the differences of the extreme, outlying group to the other cases. After entering the cluster variables as predictors and the categorized similarity measure (1= outlying, 0 = main group) as dependent, the algorithm was configured as described in Exhibit 11. For validation, a 10 fold cross-classification was chosen.

Figure 16 exhibits the results gained from this analysis. Particularly the variables “house” and charter bus contribute to explaining the variance. For example, node 4 means, that this group of travellers stays 56% or less of their nights in a house and uses the charter bus for more than 67% of their transport needs. Resubstitution and cross validation yielded a risk estimate of 1.2% (SE .06%). The solution as presented here results in 98.8% correctly classified cases (Table 40). When using only one rule (i.e. the variable “house”), the risk estimate would increase to 4.3% (SE 1.1%) with about 95.7% correctly classified cases. The improvement of classification from one rule to the final “model”, Theta (2), is 3.1% in this case. The Theta measure is calculated by subtracting the classification accuracy at one level from the accuracy of a completely developed tree. In the remaining profile of node four, not much difference could be identified to the other profiles, except that half of the group members also visited a friend or relative but did not mention this as their primary reason for travelling and 60% arrived on an inclusive tour package which can be seen linked with the usage of charter busses. None of the group members normally go on used to go on working holidays, which implies that the overall original profile is even more influenced by working holiday visitors as has been described in the segment profile above.

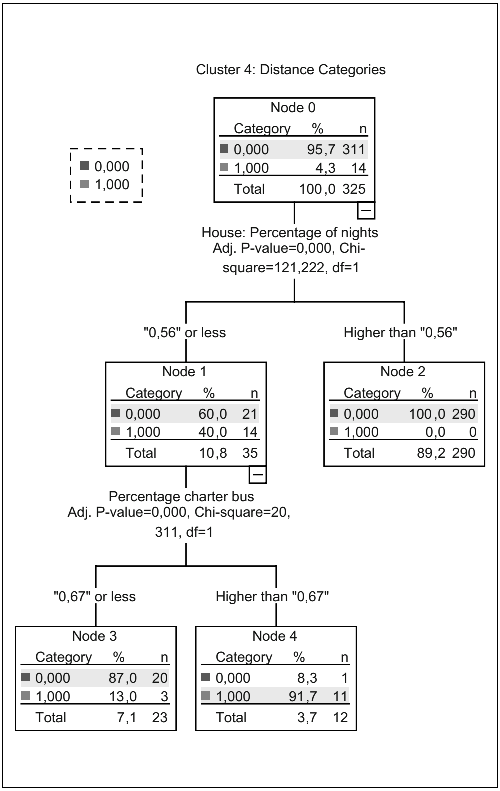


Figure 16: Graphical result cluster 4

Thus, the information derived here may help to sharpen the core profile of cluster four and help derive additional groups that may be of interest for the tourism researcher. One may also consider adapting the data collection method to account for the two different ways of using housing.

To gain a better overview, the tree derived above is entered into a decision table (see Bargeman, Joh, Timmermans and Van der Waerden 1999 for a detailed discussion on the methodology):

Prediction rules for extreme cases in cluster 4			
Node	2	3	4
House (% of nights)	> 56	≤ 56	≤ 56
Charter bus (% of transport)		≤ 67	> 67
P	100%	87%	91.7%
Predicted	Non extreme	Non extreme	Extreme
N Total	290	23	12
Statistics: 98.8% correctly classified cases, cross validation estimate .012 (SE = .006); Theta (2): 3.1%			

Table 40: Classification report cluster 4

For cluster 2, extreme groups are particularly characterized through the lower use of backpacker accommodations and no use of long distance coaches at all, as shown in Table 41.

Prediction rules for extreme cases in cluster 2			
Node	2	3	4
Backpack hostel (% of nights)	> 44%	≤ 44%	≤ 44%
Long distance coach (% of transport)		≤ 0%	> 0%
P	100%	70%	91.7%
Predicted	Non extreme	Extreme	Non extreme
N Total	311	10	24
Statistics: 98.6% correctly classified cases, cross validation estimate: .026. (SE = .009); Theta (2): 1.16 %			

Table 41: Classification report cluster 2

This underlines that the “classical” backpacker combination of “greyhound buses” and “hostels” is not necessarily valid for all backpackers.

Some even use hostels to a larger proportion (up to 44% of their overnight stays) while not using any busses at all. This can also be seen in the light of other means of transport becoming more attractive for budget travellers (e.g. airplane).

The extreme group with a score below 0.38 on the distance measure in cluster 5 was also profiled. Thereby, the variable camping ground for overnight stay acted as the main predictor (see Table 42). From node 2 to node 1, one notes that the node is even more influenced by this behaviour. In node 1, no substantial differences, except a small difference in age, could be found (i.e. M=30.3 (7.5)). In terms of accommodation, these travellers stay more often in hotels, motels or resort-type accommodation.

Prediction rules for extreme cases in cluster 5			
Node	1	2	3
Camper % of nights	≤ 30%	30 – 42	> 42
P	60%	81.8%	100%
Predicted	Extreme	Non extreme	Non Extreme
N Total	10	11	87
Statistics: 94% correctly classified cases, cross validation estimate: .074 (SE = .025); Theta (2): n.a.			

Table 42: Classification report cluster 5

Table 42 suggests that using camping is not the only means of accommodation for this group. The most extreme participants score very low (≤ 30%), which may indicate additional fragmentation of this tourist type.

To sum up, the study of global and local outliers has shed light on cruise tourists, which may represent a tourist type on its own and may become increasingly important in the future.

The investigation of extreme groups within local outlier profiles in this segmentation study has generated information for the clustering profiles. They can be interpreted in the light of tourists becoming “fragmented” themselves and, for example, staying in a tent on a campground or in the outback for some time while enjoying a five star resort for the rest of their time.

This information may be helpful in various ways: In a "classical view", it may help the "fine tuning" of average segment profiles by largely ignoring those types. For example, profiling the extreme group of cluster four, it became obvious that the main profile is very much influenced by working holiday tourists, while a group of bus tourists who also stay at houses – type accommodations for half of their overnight stays somewhat blurred the profile.

On the other hand, tourism marketers may gain the insight that seemingly contradictory combinations of transport or accommodation choices are gaining in importance. For instance, the increased use of hotels and resorts does not match the consumption pattern of the camper segment, nor does the non use of long distance coaches and a limited use of backpacker hostels fit into the classical backpacker tourist type. Nevertheless, for economic or other reasons, some may appreciate those combinations.

Chi² Automatic Interaction Detection

Chi square automatic interaction detection (CHAID) (Kass 1980) is a tree-building segmentation technique and particularly useful to predict a categorical dependent variable with the help of a large number of continuous or categorical independent variables. The major difference to other popular tree-generation algorithms, such as C 4.5 (Quinlan 1993) and CART (Breiman, Friedman, Olshen and Stone 1984), is its restriction to categorical data (Berry and Linoff 1997, p265). Nevertheless, popular CHAID software packages such as SPSS 15 and, formerly, SPSS Answertree include sophisticated options for categorizing continuous data.

The technique starts by identifying the best predictor variable for the dependent variable. This is achieved using a contingency table of the predictor and the dependent variable with the highest significance level of the chi squared statistic. In case of nominal variables, categories can be freely merged. In case of ordinal data, the ordered nature is respected. After all combinations have been calculated, they are compared against each other and the best predictor is applied. Each resulting branch is analysed the same way until a predefined stop criterion is reached (Bargeman, Joh et al. 1999). Each node can be a two-fold or multiple split. The result is a partitioning of the data into mutually exclusive, exhaustive subsets that best describe the categorized, dependent variable. As the total number of options to reduce the multidimensional table rapidly increases with an augmenting number of predictor variables, CHAID is not based on complete enumeration, but uses an alternative approach assuring satisfactory results. Two categories are merged if the significance is below a user-defined level. This process is repeated until no further merging is possible. Next, each resulting category consisting of three or more of the original categories is tested. If the most significant split of the compound category exceeds a certain threshold value, the split is implemented. These steps are repeated until no further improvement is achieved. (Bargeman, Joh et al. 1999).

Exhaustive CHAID differs to the original CHAID by delivering better results toward the strongest association between the decision and condition variables. The adjusted p value is obtained using a Bonferroni multiplier. Basically, this multiplier is determined by the number of ways a condition variable of a given type with c original condition states can be reduced to r states ($1 \leq r \leq c$). Given the adjusted p values of all conditions, the most significant condition variable (i.e. with the lowest p value) is isolated, and if the p value of this condition is less than or equal to the specified α level, then the group of observations is split according to the (merged) states of this condition variable. If no condition variable has a significant p value, the group is not split and the process terminated. For each partition of the data that has not been analyzed, the algorithm returns to the first step. The tree-growing process continues until all subgroups have either been analyzed or contain too few observations (Van Middelkoop, Borgers and Timmermans 2003). One of the earlier examples of CHAID can be taken from Armstrong and Andress (1970). They recommend decision trees if data are subject to interaction, nonlinearities or causal priorities, especially in large datasets. AID techniques are a suitable tool for group-level analysis and prediction, even though limitations have to be taken into account. (Doyle and Fenwick 1975).

Limitations of CHAID mainly arise through its "stepforward nature". The statistical tests are characterized with later effects depending on earlier ones. In comparison to regression models or analysis of variances, all effects do not necessarily fit simultaneously. Similar problems are known for stepwise regression models. (see e.g. Armstrong and Andress 1970) However, the Bonferroni multiplier corrects for the number of different ways of splitting a single predictor variable (see Biggs, de Ville and Suen 1991). The programme does not allow correcting for the number of potential splitter / predictor variables being considered. So far, there is a lack of knowledge on the implications of mixing nominal, ordinal and continuous variables in the prediction of either a nominal, ordinal or continuous dependent variable. Furthermore, the implications of different potential ways of handling missing observations are not fully investigated. Both of those limitations lack further knowledge through e.g.

Monte Carlo studies (see Huba 2003for a more extended discussion).

CHAID requires stopping criteria for splitting a given branch: This includes a predefined minimum number of cases at a leaf node and a significance level of splits. The α level for predictor choice is set at 5% following general convention in statistical analysis. As no “Hard rules” exist, recommendations from recent literature (e.g. Arentze and Timmermans 2003), are applied.

$$y = N / 100$$
$$N_0 = 6, \text{if } y < 6$$
$$N_0 = y, \text{if } y \geq 6$$

In addition, the number of equal-frequency intervals (*m*) needs to be set. The higher the value of *m*, the better the approximation to a continuous distribution. Here, *m* = 10 was chosen for those variables where no empirical experience on splits was available.

Accuracy and risk estimations are presented in misclassification tables. Furthermore, a theta measure is frequently used to describe the accuracy of a decision tree. Theta (1 column) indicates the number of correctly classified observations of a model that has only one rule. This null model represents the aggregate distribution of observation over the alternatives. In contrast, Theta (x columns) represents the number of correctly predicted cases for the final model. As a consequence, the difference between Theta(1) and Theta(x) represents the improvement in prediction accuracy (Van Middelkoop, Borgers et al. 2003).

Exhibit 11: Chi² Automatic Interaction Detection

E. Discussion and Conclusion

This chapter first summarizes the results (1), displays the limitations of this work (2), then identifies the contribution to the existing literature (3) and finally highlights major implications for researchers and practitioners (4). Section 5 identifies areas for further research.

1. Summary

This dissertation departs from the researcher's dilemma when extreme and deviant cases, also known as outliers, occur in data analysis. The researcher often needs to decide whether to retain these cases as they are, and risk the distortion of the results of an empirical study, or take special care of outliers by weighting, transforming or excluding them from the data analysis.

This dissertation has drawn on literature in different scientific disciplines, including marketing, and sheds light on the way outlying cases are used in fields such as construction, crime prevention, and innovation management. In these fields, such cases often provide much more insights than "average profiles". Moreover, the role of extreme and deviant cases in statistics literature is investigated and potential handling alternatives are delineated. Given the case that deviant and outlying observations cannot be attributed to errors in data collection or handling, they are considered as part of the phenomenon under study. The available options for handling them include accommodation, rejection (i.e. deleting them), identification for separate study, incorporation into a revised model, or further experimentation (reworking the whole study including data collection).

As research paradigms strongly influence or even determine research practices, all outlier-handling alternatives were discussed in the light of positivist, critical rationalist and realist research paradigms, as those allow empirical data and statistics as base for scientific knowledge. Thereby, deleting and reweighting outliers could be attributed to a logical empiricist paradigm following Wittgenstein's verificationist approach, while "identification for separate study", "sensitivity analysis" and "incorporation" were found to be compatible with a critical realist paradigm. Accommodation per se (i.e. the adaptation of the measurement scale for the analysis procedure) could not be identified as compatible with any of the paradigms under study.

To estimate the importance of the phenomenon per se and the related handling alternatives, a publication audit on premier marketing publications was conducted. This audit identified 115 articles reporting outliers in their data analysis (n = 3,918). In these articles, 50% deleted these cases while 13% transformed extreme values and thus accommodated them. Statistical rules of thumb (50%) were the most frequent arguments for outlier treatment, even though the reasoning of such rules could not be found in statistics literature. In 26% of the cases, no reasoning at all was found. Both treatments are considered problematic from a critical realist point of view. The publication audit also detected an additional alternative for outlier handling, “sensitivity analysis”. In sensitivity analysis, results are calculated twice, once including and once excluding outliers and then compared.

To study the consequences of deleting cases on the results, an empirical study was conducted. Becken and Gnoth’s study (2004) on tourist types (i.e. market segments) published in the journal “Tourism Management” served as a benchmark and was replicated. In the original study, the authors used censoring rules to “presegment” data from the international visitor survey (IVS) to New Zealand, but did not report the effect of excluding cases on the results. In the replication study, this dissertation first analysed data from a similar survey on tourists to Australia and then repeated the procedure using the same censoring rules.

The comparison of the results showed that homogeneity of the data was substantially increased through the application of censoring rules. The cluster analysis per se led to four profiles roughly similar to the findings of Becken and Gnoth (2004), whereas one completely different group profile emerged. Interestingly, the groups found in the censored sample largely converged with the results of the reference study as well as with documented tourist types for Australia. The uncensored analysis identified a group that, at first sight, substantially deviated from current understandings of tourist types in Australia. Nevertheless, this group could also be validated via external data and detailed within case analysis and thus cannot be considered an error driven. In the clustering procedure of the uncensored data, one outlying group emerged which was not present in the censored data. Their profile unvealed them as cruise ship tourists. While cruise ships are important sources of visitors for sea side tourist destinations, Australia has not taken much advantage of them compared to other destinations. Recent studies unveiled the potential benefits for Tourism Australia, such as additional sources of revenues from the globally growing cruise tourism industry. Using the censored approach, this tourist type would have remained undiscovered.

Overall, the exclusion of outlying or deviant cases can be seen in the light of the confirmation bias existing in many areas of marketing research. That means that the researcher seeks to confirm his (or the research field's) a priori hypotheses. The confirmation bias has been investigated in the history of science and shown to delay progress in research. In psychology, for example, the confirmation bias led to 25 years of delay in the investigation of the sleeper effect. In the case of the ozone hole, the exclusion of outliers led to a delay of almost 10 years until the substantial decline of thickness was unveiled. From a philosophy of science point of view, confirmation bias is also often related to the verificationist view promoted by the Vienna Circle and logical empiricists thought.

2. Limitations

Besides these insights, some limitations are inherent to this dissertation:

The publication audit studies the behaviour of researchers by inspecting their published papers. As with any publication audit or a meta analysis, the "file drawer problem" is also inherent to this audit: A large proportion of papers is not published and remains in the "researcher's file drawer". Even though top tier marketing journals only represent a small portion of research in this area, it reflects the "most admired" view in the discipline, which is probably even more interesting at this point.

The publication audit identified 212 articles for detailed coding through a semi-automatic search procedure using two different databases and nine popular terms. Thus, publications labelling outlying cases in a different way as those used in the search library may have been left out. On the other hand, the chances are relatively small as the search library has been compiled after also studying the phenomenon in other disciplines, such as statistics and data-mining. Additionally, studies that do not declare that they have undertaken outlier handling are likely to present more appealing results and thus increase the likelihood of being published under current conditions. They could not be identified through the data collection procedure used. Even worse, their results may influence the direction of a scientific field.

A publication audit, nevertheless, does not reflect whether the outlier handling and rejection is a problem on the author's or reviewer's side. However, they may also be overlapping. To initiate countermeasures, it would be valuable to know which role (i.e. authors, reviewers, or influencers) holds and promotes these practices.

Related to the replication of Becken and Gnoth's (2004) article, several limitations apply which have also been stated as such in the sections concerned. The data collection instrument (IVS) for Australia differs somewhat from the one used in New Zealand. For instance, activity variables have been recorded in a different way. They are incompatible with the research approach and cannot be summarized in a meaningful way. Thus, they have not been included in the cluster analysis of this study, which represents a deviation from the original approach. As Becken and Gnoth (2004) argue that those variables did not hold any discriminatory potential between groups, this deviance was neglected.

Additionally, the New Zealand IVS instrument also covered business travellers (6.9%), conference participants (1.5%) and educational travellers in their dataset (3.2%). Even though this may have influenced the analysis, the simulation was pursued on highly similar variables and the censoring effect was evaluated and interpreted only in relationship to the Australian dataset. Furthermore, the validation was conducted by using data external to this study and showed a realistic and sound result.

While the reference study used kilometres as the unit for transport modes, the present study used the number of times a specific means of transport was chosen as kilometres were not available. This detail may not carry as much weight as the correlation measure used for clustering does not take into account the actual size but only the pattern. In the reference study, the elimination of outliers has not been stated as a goal, but a priori censoring rules have been used instead, which is different to many other publications.

The philosophical positions addressed in this work refer only to the main streams in this field. Wherever possible, substantial care was taken to use these positions and paradigms in a precise way. Given that some of the paradigms have rather fuzzy boundaries, the evaluation in this dissertation refers only to the cited individuals and their views.

3. *Contribution to the literature*

Overall, this dissertation deals with a problem already discussed by Bernoulli (1777) and proposes a different approach to address the issue of whether extreme cases are to be considered as "statistical noise or valuable information". Thereby, the epistemological position of the researcher (i.e. their "world view") plays an important role in the decision on handling alternatives: While (logical) positivists largely consider extreme and outlying cases as errors, critical realists are interested in these cases and their nature. Thus, depending on a researcher's own position, this dissertation discusses the relevant handling alternatives. In order to quantify the issue, a publication audit on the marketing literature shows that the option "outlier deletion without reasoning", related to logical empiricism, dominates the field.

Excluding or altering parts of the sample also has consequences for the use of inference statistics where censoring or modifying a sample destroys its probability characteristics and thus the possibility of making generalizations. In a strict interpretation, results from censored samples do not represent the population in which the researcher is interested, but only the sample used. To illustrate the consequences of the outlier deletion alternative, the emipircal study of this dissertation investigated the consequences of censoring on the results by replicating a study published in the journal "Tourism Management" and concluded that substantial differences can arise in the results, which may mislead other researchers and practitioners. If in doubt about whether outliers are error laden or not, this dissertation proposes calculation and reporting of a sensitivity analysis instead of publishing confirmation-biased and misleading results and shows this approach with the dataset at hand.

Finally, extreme and outlying cases showed the potential to create new scientific knowledge: In the empirical study, new and distinct groups of consumers could be identified that could not be classified with current knowledge and which are discussed in other parts of the literature as an economically-important profile.

In a more detailed way, this work adds and completes current knowledge of the field in different areas:

First, this dissertation empirically underlines what some researchers in social science suspected: Outliers are often deleted or averaged out (e.g. Julander and Söderlund 2002), often by using statistical rules of thumb. Arguments, affected methods and differences between journals have been outlined. This finding complements the survey study by Orr, Sacket et al. (1991) where about 4% of the surveyed authors were found to commit outlier deletion without any further reasoning, 29% include all data points and 67% exclude cases only if a serious rationale exists. Warnings by senior authors (e.g. Finney 2006), and somewhat unspecific recommendations in the “APA” publication guidelines (Wilkinson 1999), can be found in the marketing literature and relate to this topic. In Section C.3 of this dissertation, the existence of a confirmation bias could be unveiled in other phases of the research process. This includes the widespread use of null hypothesis significance testing, where the null hypothesis is specified in a way that it is almost always rejected (Gigerenzer, Krauss et al. 2004). Overall, one can state that the decision on how to handle outliers is frequently taken in a subjective way (see e.g. Gummesson 2003 for a discussion on subjectivity in research), not taking into account the consequences of publishing confirmation-biased results, which may even be favoured by the academic publishing and incentive system.

Second, this dissertation discussed different research paradigms i.e. the assumptions of the researcher, which influence every research project. This is particularly important, as research paradigms frequently limit the discovery of scientific findings or influence the “reality” which can be discovered. This discussion unveiled that the measures of outlier treatment used by the researcher depend on the research paradigm to which they subscribe. Logical positivists, for example, tolerate the deletion of extreme and deviant cases, as only confirming evidence is important, whereas disconfirmation is sometimes even called error. Critical rationalists in contrast regard them as important for theoretical progress. This finding also supports the existence of different views on outliers by the community. In a relativist or constructivist notion, outlying data may need to be reinterpreted in the circumstances where it emerged. Approaches compatible to a critical realist view of science are discussed as, according to Hunt (1994), all research in marketing should be compatible with the critical realist paradigm. The paradigm also influences whether researchers perceive outliers as statistical noise or valuable information.

Third, this study empirically investigated the effect of censoring on the results of a published study. While replication per se is important for scientific progress (e.g. Armstrong 2003), this study also shows the effect on sample heterogeneity and the subsequent results. Similar studies have been conducted by Mullen, Milne et al. (1995) using SEM modelling in marketing and by Orr, Sackett et al. (1991) in psychology. To the author's knowledge, the present study is the first using a cluster analysis procedure. The study additionally profiles outlying cases and shows that they can be very "realistic".

Forth, this study illustrates how to derive additional insights in a cluster analysis study. While Julander and Söderlund (2002) have shown this for regression analysis, the investigation techniques for global and local outliers in a cluster analysis study are identified and tested here. As many forms of cluster analysis are sensitive to extreme and outlying values, the exclusion of small parts of the sample can result in completely different results and thus other conclusions. Additionally, these extreme groups can, as in the present case, unveil small, homogeneous segments representing upcoming, unknown segments or known segments loosing importance.

Fifth, this dissertation critically adds to the domain of tourism research, where Becken and Gnoth's (2004) study could be replicated. Additionally, the widespread practice of "presegmentation" (i.e. censoring) is criticised as the effects of this pre-segmentation are not reported. In this study, the outlying group of ship tourists was excluded while they are considered an economically-relevant group in other publications.

Sixth, this study showed that the analysis of outlying groups might also help to sharpen the core profile of derived market segments. In segmentation analysis, most of the heterogeneity is frequently not addressed (Franke, Reisinger et al. 2007). That means that a few groups are usually formed to decrease the variability in the dataset, but the within-group variance is still substantial. Investigating the sources of this heterogeneity may help to improve, for example, the positioning of a product to be delivered.

4. Implications

This dissertation has shown that censoring data and neglecting outliers may have substantial consequences for the results of a study. On the one hand, outlier data can include additional findings that are then simply overlooked, while censoring violates assumptions for statistical inference. On the other hand, dealing with outliers uncritically can also mislead other researchers who build on such a work and consequently waste effort and resources. One of the reasons for this phenomenon can be seen in the way academic work is evaluated and promoted: Publishing irregular or negative results proved to be difficult to impossible. However, it may be worthwhile as it prevents other researchers from taking the wrong, confirmation-biased path. Adapting incentive systems and publication outlet policies may be the first steps to work in this direction.

In case a critical realist researcher faces outlying and deviant observations, they are advised to first inspect the cases and to determine whether the values occur naturally or are error laden. In the former case, one of the following options may be advisable: Integrating outlying observations in an improved model or the reasoned change of the reference distribution may be an option. Depending on the type of analysis, also a separate study of outliers can be worthwhile. The research undertaken here suggests that the deliberate investigation of outlying values may give additional theoretical insights or point to small phenomena that may be interesting for future study. This is in line with a stream of "non confirmation-biased" modifications to the current use of statistical tools such, as the reporting of power and effect sizes, and the use of competing hypotheses, just to name a two. In the case of error, data re-collection, correction of errors or exclusion may be worthwhile considering. If unsure, analysing the sensitivity of the results by calculating a full and a censored version can be a solution.

In his studies on the "tools-to-theories" heuristic, Gigerenzer found that after researchers adopted statistical test procedures and their implementations in statistical software, these procedures evolved as their own theory for the mind. Their use in current projects became a "ritual"; the quest for the most suitable procedure was suspended. From an instrumentalist point of view, the critical investigation of extreme and deviant cases, implemented in research practices (e.g. checklists, textbooks) and software applications, could help to derive additional insights and help to influence undesired paradigmatic influences.

As this dissertation exhibits a methodological focus, managerial implications strongly resemble those mentioned for researchers. The applications of the results are manifold and support existing practices in some areas.

In market research, agencies or consulting companies often analyse data for a client company. Thereby, clean data and clear analyses are often considered as a sign for quality of work and a pre-requisite for future collaborations. Thus, both clients and researchers should take care as to whether the data are censored or weighted for analysis. Many studies may profit from having the possibility to re-interview certain extreme participants as they may represent economically-interesting market segments, such as the cruise ship group in this dissertation's study. This suggestion may seem difficult in terms of anonymity, but trustworthy intermediaries, such as universities or research institutes may help to provide a workaround solution. Studies undertaken on the entire population (such as the customer segmentation here) may include indicators for new segments or micro trends. The critical re-investigation and discussion of past research data may be worthwhile to unveil insights, such as trend indicators.

Similar implications arise for behavioural data, which – due to the nature of transaction – is largely available to, for example, telecom or financial services providers. Extreme trading or communication patterns may, for example, exhibit arising needs, but also problems in billing or the offer. Similarly, intruders into computer networks exhibit extreme behavioural patterns and so can be detected. In both applications, a very narrow focus on business questions can result in substantial threats or lost opportunities. Researching these cases using different, more in-depth, methods may lead to fruitful insights. Section B.1 discussed further applications for extreme cases.

5. Areas for further research

In this dissertation, confirmation bias was found to have a substantial influence on the results of a study. A study checking for multiple symptoms at a time may reveal which parts of a study are particularly prone to be affected. An analysis on the "social network" of confirmation-biased researchers may additionally help to identify schools of thought promoting confirmation-biased research. Symptoms or indicators that may be used include the use of dominant hypotheses, selective literature reviews and the handling of outliers for example.

Related to the publication audit of this dissertation: It can be extended to include second and third tier marketing journals and then allow inference on the publication policies and their effectiveness, as well as provide data for an interpretation in a timeline way. Questions may include whether or not the way outlier handling has changed over time. This could be identified by, for example, administering a submission study including critical realist and confirmationist practices (see Armstrong and Hubbard 1991).

With respect to the limitations shown in the replication of Becken and Gnoth's (2004) study: A comparison to a similar analysis on the same data a few years later may provide information on the predictive value of the extreme group found. Additionally, the collection of more detailed itinerary data may allow a closer replication of the original study.

This study has investigated the effect of censoring on a cluster analysis. While clustering methods are known to be sensitive to changes in the structures of the data, a replication of this study in another context and with different censoring levels would be desirable. Studies with other methods, such as popular structural equation modelling, may also provide fruitful areas of research.

In the area of marketing research, the survey method is still frequently used. A comparative analysis of surveys facing the "outlier problem" may allow drawing conclusions on similar patterns or traits inherent to outlying cases. Similar to the studies on the involvement levels of survey participants versus non participants (e.g.Wijnen, Vermeir and Van Kenhove 2007), this comparison would allow for speculation and inference on "outlier-generating mechanisms", if they exist.

This dissertation focussed on the marketing research using statistics and data analysis to study marketing phenomena. A comparison with other fields, and particularly those nearer to mathematics and statistics, may be interesting to unveil whether the issues presented in this dissertation are limited to the marketing domain. As confirmation biased results have also been identified in other domains, this study may contribute to gaining a better idea of the "big picture" of the problem in science.

Besides these areas where additional research looks very promising, this dissertation has already provided evidence that extreme and outlying cases can contribute valuable aspects to scientific knowledge and are not necessarily error driven per se. This learning questions the dominant logic in marketing research of deleting such cases without further reasoning. If this logic is applied, results can be misleading for the future development of scientific knowledge in the area. The empirical study has shown that substantial differences in the results can arise whether extreme and outlying cases are excluded or not. In this study, these cases evoked potentially interesting insights such as an upcoming new segment.

This dissertations therefore concludes with the recommendation to first look for the source or generating mechanism of such extreme and outlying cases and then select the treatment option carefully and with reason. In the case that the source or the mechanism cannot be identified (e.g. because the relevant interviewee is no more available), the researcher could for instance decide to report the result of a sensitivity analysis along with the findings. This would allow other researchers to build on these findings. Additionally, the in depth study of these cases may be very valuable as a study per se or as inspiration for new projects.

APPENDIX

Appendix 1: Consumption profiles test statistics

Full Data

Cluster Variables Test data								
	Group 1	Group 2	Group 3	Group 4	Group 5			
	M % (SD)	M % (SD)	M % (SD)	M % (SD)	M % (SD)	F (df)	p =	Effect size
Hotel and Resort	3.2 (9.4)	2.6 (8.2)	89.5 (23.2)	2.7 (8.6)	7.3 (10.6)	8437.82 (4)	0.00	0.931
Motel	2.2 (10.6)	1.9 (9.1)	4.7 (17.8)	5.1 (18.8)	5.1 (14.2)	10.38 (4)	0.00	0.089
Backpacker	1.1 (5.7)	78.6 (23.8)	0.4 (3.1)	2.8 (7.8)	6.4 (13.2)	7709.6 (4)	0.00	0.925
House	0.9 (5.8)	2 (8.3)	0.5 (4.1)	85.4 (24)	2.6 (11.2)	8439.34 (4)	0.00	0.931
Farm and B&B	0.3 (4.1)	0.6 (5.2)	0.7 (5.9)	0.3 (5.5)	0.3 (1.8)	1.54 (4)	0.19	0.034
VFR, Home	91 (18.6)	7.8 (14.2)	2.2 (8.8)	2.6 (7.9)	6.7 (14.2)	12133.79 (4)	0.00	0.951
Camper	1 (5.6)	3.9 (8.6)	0.6 (4)	0.7 (3.7)	69.6 (27.3)	3019.38 (4)	0.00	0.837
Boat	0.2 (2.9)	2.1 (11.3)	1.2 (7.1)	0.2 (2.9)	1.8 (10)	13.9 (4)	0.00	0.103
Car transport	60.2 (44.7)	6.6 (18.2)	3.6 (16.7)	12 (29.2)	7.9 (21.6)	921.48 (4)	0.00	0.645
Rental car transport	3.6 (16.8)	7 (21.1)	9.2 (25.6)	16.5 (34.7)	20.8 (36.4)	39.8 (4)	0.00	0.173
Van transport	0 (0.4)	0.4 (3)	0.1 (2.5)	0.2 (2.6)	42.8 (38.9)	1370.91 (4)	0.00	0.718
Taxi or	5.6	2.7	19.3	6.9 (23)	3.3	15.49	0.00	0.109

chauffeur transport	(21.6)	(12.7)	(27.6)		(10.8)	(4)		
Plane transport	21.6 (37)	16.1 (25.7)	36 (37.3)	16.6 (32.8)	8.9 (12.9)	68.62 (4)	0.00	0.225
Train (LD) tranport	1.3 (8.8)	2.1 (8.9)	0.6 (5.4)	1.4 (8.8)	0.1 (0.9)	5.16 (4)	0.00	0.063
Coach (LD) transport	2.6 (12.7)	31.9 (33.3)	2.1 (10.9)	8.5 (23.1)	1.6 (6.2)	314.16 (4)	0.00	0.442
Ship, boat transport	0.6 (4.7)	4.9 (12.6)	2 (9.4)	1 (4.6)	2.7 (10.4)	26.75 (4)	0.00	0.142
Public transport	2 (12.6)	12 (28.1)	2.3 (13.8)	2.4 (13.6)	2 (10.1)	36.77 (4)	0.00	0.166
Shuttle transport	0.5 (5.1)	4.1 (14.6)	8.6 (25.1)	1.6 (10.4)	0.7 (2.7)	63.67 (4)	0.00	0.217
Charter bus transport	1.8 (12)	9.5 (23.3)	24.7 (36.4)	32.1 (45.2)	2.5 (11.4)	223.15 (4)	0.00	0,384

Censored Data

Cluster Variables Test data								
	Group 1	Group 2	Group 3	Group 4	Group 5			
	M % (SD)	M % (SD)	M % (SD)	M % (SD)	M % (SD)	F (df)	p =	Effect size
Hotel and Resort	9.5 (15.3)	87.9 (25.5)	15.2 (19)	3.9 (10.8)	9.6 (13.5)	2087.79 (4)	0.00	0.881
Motel	2.8 (9.8)	3.5 (13.4)	26.3 (29.5)	0.5 (2.4)	2.8 (6.2)	130.61 (4)	0.00	0.423
Backpacker	4.4 (13.7)	0.5 (3.7)	9.1 (23.3)	78.5 (22.5)	7.1 (14.6)	1457.32 (4)	0.00	0.842
House	3 (12.4)	3.5 (15.7)	14.5 (28.6)	3 (12.1)	2.5 (10.8)	23.09 (4)	0.00	0.192
Farm and BandB	0.4 (3.9)	0.6 (5.4)	3.6 (12.4)	0.3 (1.8)	0.1 (0.9)	13.73 (4)	0.00	0.150
VFR, Home	76.7 (24.6)	3.2 (10.1)	20 (29.9)	9.2 (15.4)	8.9 (16.9)	1928.66 (4)	0.00	0.873
Camper	2.9 (10.7)	0.5 (3.5)	10.3 (24.7)	4.2 (9.3)	68.6 (26.4)	601.89 (4)	0.00	0.708
Boat	0.1 (0.6)	0 (0)	0 (0.3)	0.2 (1.3)	0.3 (1.2)	11.27 (4)	0.00	0.136
Car transport	49.3 (38.9)	2.9 (11.5)	2.7 (7.1)	2.2 (6)	2.2 (6.6)	505.33 (4)	0.00	0.676
Rental car transport	1.3 (7.2)	7.2 (21.3)	83.4 (16.2)	2.2 (8)	2.2 (5.9)	1086.22 (4)	0.00	0.803
Van transport	0.1 (1.3)	0.1 (2)	0.1 (0.9)	0.4 (3.7)	62.8 (32.5)	1954.31 (4)	0.00	0.875
Taxi or chauffeur transport	2.5 (9.8)	4.9 (13.9)	2 (5.4)	2.2 (7.2)	2.2 (5)	7.56 (4)	0.00	0.111
Plane transport	30.5 (36.2)	52.7 (30.4)	7 (12)	20.1 (26.9)	10.5 (11.5)	156.9 (4)	0.00	0.455
Train (LD) tranport	3.6 (14.1)	1 (7.2)	0.6 (3.3)	2.4 (10.1)	0.3 (1.4)	9.76 (4)	0.00	0.126

Coach (LD) transport	6.3 (18.1)	2.9 (12.3)	0.3 (2.5)	47.3 (32.8)	1.6 (6.5)	288.28 (4)	0.00	0.570
Ship, boat transport	0.1 (1)	0.1 (0.6)	0.2 (1.2)	0.5 (1.8)	0.6 (1.9)	11.86 (4)	0.00	0.139
Public transport	2.2 (10.2)	1.1 (7.4)	0.9 (3.4)	4.7 (10.8)	0.8 (2.7)	7.99 (4)	0.00	0.115
Shuttle transport	0.9 (5.9)	4 (12.9)	1.4 (6.2)	2.7 (7.7)	1.1 (3.4)	12.68 (4)	0.00	0.144
Charter bus transport	2.3 (11.6)	22.4 (28.2)	0.9 (4)	11.3 (25.9)	1.9 (7.2)	119.62 (4)	0.00	0.408

Censored Data

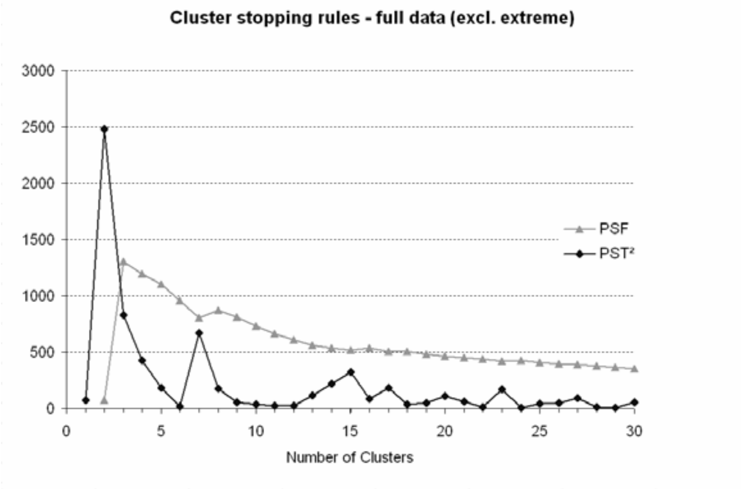
Descriptive Variables – Test Data								
Variables						Chi² (df)	p=	Cramer V
Country						1202.56 (252)	0.000	0.350
First time/repeat						319.45 (4)	0.000	0.360
Group tourist						613.78 (4)	0.000	0.510
Travel party						599.47 (16)	0.000	0.250
Agegroup						314.83 (44)	0.000	0.180
Sex						11.3 (4)	0.023	0.070
	Group 1	Group 2	Group 3	Group 4	Group 5			
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	F (df)	p=	Effect size
Total expenditure	2420 (3448)	2880 (7346)	4402 (4975)	3508 (2695)	4898 (4314)	6.7 (4)	0.000	0.107
expenditure p.d.	111 (153)	326 (807)	260 (427)	108 (60)	144 (89)	17.89 (4)	0.000	0.173

Length of stay	29.9 (31.1)	10.7 (10.8)	21.3 (17.2)	40.5 (33.8)	34.6 (28.5)	132 (4)	0.000	0.425

Full Data

Descriptive Variables – Test Data								
Variables						Chi² (df)	p=	Cramer V
Country						2061.37 (308)	0.00	0.316
First time/repeat						670.79 (4)	0.00	0.360
Group tourist						1079.4 (4)	0.00	0.457
Travel party						1201.3 (16)	0.00	0.241
Agegroup						647.89 (44)	0.00	0.177
Sex						56.89 (4)	0.00	0.105
	Group 1	Group 2	Group 3	Group 4	Group 5			
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	F (df)	p=	Effect size
Total expenditure	1792 (2788)	3706 (3972)	2547 (5753)	5184 (8498)	5428 (5130)	49.98 (4)	0.00	0.195
expenditure p.d.	128 (259)	116 (100)	350 (710)	198 (310)	164 (110)	58.08 (4)	0.00	0.210
Length of stay	24.2 (37.7)	52.9 (73.2)	8.8 (8.4)	63.0 (102.0)	39.4 (45.2)	189.97 (4)	0.00	0.358

Appendix 2



Tourist type	N / %	Travel choice variables	Descriptive variables for each tourist type
1) VFR	2217 / 42.9%	60% car 22% airplane 90% family or friend's home	21% NZL, 14% UK, 7% JAP, 7% USA 71% return visitors, 99% no group tourists, 69% unaccompanied traveller Age: 40 yrs, 61% female Total expenditure: 1911 $, expenditure per day:128 $ Stay: 26d
2) Back-packer	277 / 5.3%	34%coach 18% airplane 83% backpacker hostel	26% UK, 10% JAP, 9 % GER 76% first time visitors, 97% no group tourists, 66% unaccompanied traveller Age: 26 yrs, 52% male Total expenditure: 3433 $, expenditure per day: 114 $ Stay: 43d
3) Comfort tourist	2359 / 45.6%	33% airplane 24%charter bus 18% taxi or shuttle 84% hotel or resort	35% JAP, 9% USA, 9% NZL 62% first time visitors, 34%group tourists,47% travelling couple Age: 37 yrs, 51% female Total expenditure: 2607 $, expenditure per day:341 $ Stay: 10 d
4) Housestay & charterbus tourist	238 / 4.6%	39% charter bus 24% airplane 15% Long Distance Coach 85% house	35% NZL, 11% SIG, 11% UK, 10% JAP 42% first time visitors, 97% no group tourists,35% travelling couple, 37% unaccompanied Age: 40 yrs, 59% female Total expenditure: 5420 $, expenditure per day: 186$ Stay: 79 d
5) Camper	77 / 1.5%	59% van 66% camper	35% GER, 20% UK, 12%NED, 10% SUI 64%first time visitors, 96% no group tourists, 49% travelling couple Age: 35 yrs, 52% male Total expenditure: 5775 $, expenditure per day: 145 $ Stay: 39 d

F. Literature

Abraham, B. and G. E. P. Box (1978). "Linear Models and Spurious Observations." Applied Statistics **27**: 120 - 130.

Aggarwal, C. and P. Yu (2001). Outlier detection for high dimensional data. SIGMOD'01, Santa Barbara, CA.

Agnew, R. (1992). "Foundation for a general strain theory of crime and delinquency." Criminology **30**(47 - 87).

Agrawal, J. and W. A. Kamakura (1995). "The Economic Worth of Celebrity Endorsers: An Event Study Analysis." Journal of Marketing **59**(3): 56 - 62.

Airy, G. B. (1856). "Letter From Professor Airy, Astronomer Royal, to the Editor." Astronomical Journal **90**: 137 - 138.

Ajzen, I. (1988). Attitudes, personality and behaviour. Milton Keynes, Open University Press.

Ajzen, I. (1991). "The theory of planned behaviour." Organizational Behaviour and Human Decision Processes **50**: 179 - 211.

Ajzen, I. and B. L. Driver (1992). "Application of the theory of planned behaviour to leisure choice." Journal of Leisure Research **24**(3): 207 - 224.

Ajzen, I. and M. Fishbein (1980). Understanding attitudes and predicting social behaviour. Englewood Cliffs, NJ, Prentice Hall.

Aldenderfer, M. and R. K. Blashfield (1984). Cluster Analysis. Beverly Hills, Sage.

Allenby, G. M. and J. Ginter (1995). "Using Extremes to Design Products and Segment Markets." Journal of Marketing Research **32**(November): 392 - 403.

Amine, L. S. (1993). "Linking consumer behaviour constructs to international marketing strategy a comment on Wills, Samli, and Jacobs and an extension." Journal of the Academy of Marketing Science **21**(1): 71 - 77.

Anderson, P. F. (1983). "Marketing, Scientific Progress, and Scientific Method." Journal of Marketing **47**(4): 18-31.

Anderson, P. F. (1986). "On method in consumer research: A critical relativist perspective." Journal of Consumer Research **13**: 155 - 173.

Andrews, D. F., P. Bickel, F. Hampel, P. J. Huber, W. Rogers and J. W. Tukey (1972). Robust Estimate of Location, Princeton University Press.

Andrews, D. F. and D. Pregibon (1978). "Finding Outliers That Matter." Journal of the Royal Statistical Society Series B-Methodological **40**(1): 85-93.

Angeles, P. A. (1992). The Harper Collins Dictionary of Philosophy. New York, Harper Collins Publishers.

Angiulli, F. and C. Pizzuti (2002). Fast outlier detection in high dimensional spaces. Proceedings of the 6th European Conferece on Principles of Data Mining and Knowledge Discovery.

Anscombe, F. J. (1960). "Rejection of Outliers." Technometrics **21**: 123 - 147.

Anscombe, F. J. and B. A. Barron (1966). "Treatment of Outliers in Samples of Size 3." Journal of Research of the National Bureau of Standards Section B-Mathematical Sciences **B 70**(2): 141-&.

Arabie, P. and L. Hubert (1994). Cluster Analysis in Marketing Research. Advanced Methods of Marketing Research. R. P. Bagozzi. Cambridge, MA, Blackwell.

Arce-Ferrer, A. J. (2006). "An Investigation Into the Factors Influencing Extreme - Response Style." Educational and Psychological Measurement **66**(3): 374 - 392.

Arentze, T. A. and H. J. P. Timmermans (2003). "Measuring the Goodness-of-Fit of Decision Tree Models of Discrete and Continuous Activity-Travel Choice - Methods and Empirical Illustration." Journal of Geographical systems **4**: 1-22.

Arimond, G. and S. Lethlean (1996). "Profit Center Analysis within Private Campgrounds." Journal of Travel Research **34**(4): 52 - 58.

Armitage, C. J. and M. Conner (2001). "Efficacy of the Theory of Planned Behaviour: A meta-analytic review." British Journal of Social Psychology **40**: 471 - 499.

Armstrong, G. (2007). "Significance Tests Harm Progress in Forecasting." International Journal of Forecasting **23**: 321 - 327.

Armstrong, G. and J. G. Andress (1970). "Exploratory Analysis: Trees vs. Regression." Journal of Marketing Research **VII**(Novermber): 487 - 492.

Armstrong, J. S. (2003). "Discovery and communication of important marketing findings - Evidence and proposals." Journal of Business Research **56**(1): 69-84.

Armstrong, J. S., R. J. Brodie and A. G. Parsons (2001). "Hypotheses in marketing science: Literature review and publication audit." Marketing Letters **12**(2): 171-187.

Armstrong, J. S. and R. Hubbard (1991). "Does the Need for Agreement among Reviewers Inhibit the Publication of Controversial Findings?" Behavioral and Brain Sciences **14**(1): 136-136.

Arndt, J. (1985). "On making marketing science more scientific: the role of observations, paradigms, metaphors and puzzle solving." Journal of Marketing **49**(Summer): 11 - 23.

Arning, A., R. Agrawal and P. Raghavan (1996). "A Linear Method for Deviation Detection in Large Databases." KDD'96: 164 - 169.

Arrigo, B. A. and T. J. Bernard (1997). "Postmodern Criminology in Relation to Radical and Conflict Criminology." Critical Criminology **8**(2): 38-60.

Assael, H. (1987). Consumer behavior and marketing action. Boston, MA, PWS-Kent Publ.Co.

Ateljevic, I. and S. Doorne (2001). Nowhere Left to Run: A study of value boundaries and segmentation within the backpacker market of New Zealand. Consumer Psychology of Tourism, Hospitality and Leisure. G. I. Crouch, J. A. Mazanec, J. B. R. Ritchie and A. G. Woodside, CABI publishing. **2**.

Atkinson, A. C. (1981). "Robustness, Transformations and Two Graphical Displays for Outlying and Influential Observations in Regression." Biometrika **68**: 13 - 20.

Atkinson, A. C. (1982). "Regression Diagnostics, Transformations and Constructed Variables." Journal of the Royal Statistical Society Series B-Methodological **44**(1): 1-36.

Atkinson, A. C. (1986). "Masking Unmasked." Biometrika **73**(3): 533-541.

Atkinson, D. R., M. J. Furlong and B. E. Wampold (1982). "Statistical Significance, Reviewer Evaluations, and the Scientific Process - Is There a (Statistically) Significant Relationship." Journal of Counselling Psychology **29**(2): 189-194.

Ayer, A. J. (1952). Language, Truth and Logic. New York, Dover Publications.

Bachman, J. G. and P. M. Omalley (1984). "Yea-Saying, Nay-Saying, and Going to Extremes - Black-White Differences in Response Styles." Public Opinion Quarterly **48**(2): 491-509.

Bacon-Shone, J. and W. K. Fung (1987). "A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data." Applied Statistics **36**(2): 153 - 182.

Bagozzi, R. P., H. Baumgartner and Y. J. Yi (1992). "State Versus Action Orientation and the Theory of Reasoned Action - an Application to Coupon Usage." Journal of Consumer Research **18**(4): 505-518.

Ball, P. (2004). Critical Mass: How One Thing leads to Another. New York, Farrar, Straus and Giroux.

Baloglu, S. and K. W. McCleary (1999). "U.S. International Pleasure Travellers' Images of Four Mediterranean Destinations: A Comparison of Visitors and Nonvisitors." Journal of Travel Research **38**(2): 144 - 152.

Barabási, A. L., L. Albert, H. Jeong and G. Bianconi (2000). "Response to: Power - Law Distribution of the World-Wide Web." Science **287**(2115a).

Barabási, A. L. and E. Bonabeau (2003). "Scale - Free Networks." Scientific American **288**(May): 60 - 69.

Barett, D. (1997). The paradox process: creative business solutions, where you least expect to find them. New York, AMACOM.

Bargeman, B., C.-H. Joh, H. J. P. Timmermans and P. Van der Waerden (1999). "Correlates of Tourist Vacation behaviour: A Combination of CHAID and Loglinear Logit Analysis." Tourism Analysis **4**: 83 - 93.

Barnett, V. (1978). "The Study of Outliers: Purpose and Model." Applied Statistics **27**(3): 242 - 250.

Barnett, V. (1979). "Some Outlier Tests for Multivariate Samples." South African Statistical Journal **13**: 29 - 52.

Barnett, V. (1983a). Principles and Methods for Handling Outliers in Data Sets. Statistical Methods and the Improvement of Data Quality. T. Wright. Orlando, Fl, Academic Press Inc.

Barnett, V. (1983b). "Reduced Distance Measures and Transformations in Processing Multivariate Outliers." Australian Journal of Statistics **25**(1): 64 - 75.

Barnett, V. and T. Lewis (1994). Outliers in Statistical Data. Chichester, John Wiley & Sons Ltd.

Bartels, R. (1951). "Can Marketing Be a Science)." Journal of Marketing **15**(3): 319 - 328.

Basawa, I. V., R. M. Huggins and R. G. Staudte (1985). "Robust-Tests for Time-Series with an Application to 1st-Order Auto-Regressive Processes." Biometrika **72**(3): 559-571.

Bass, F. M., D. J. Tigert and R. T. Lonsdale (1968). "Market Segmentation: Group Versus Individual Behavior." Journal of Marketing Research **5**(August): 264 - 270.

Battaglia, F. and L. Orfei (2005). "Outlier detection and estimation in nonlinear time series." Journal of Time Series Analysis **26**(1): 107-121.

Batteye, R. and T. Suridge (2003). Annual Results of the International Visitor Survey 1999 - 2002. Canberra, Bureau of Tourism Research.

Baumann, R. (2000). Marktsegmentierung in Sozial- und Wirtschaftswissenschaften. Institut für Tourismus und Freizeitwirtschaft. Wien, Wirtschaftsuniversität Wien**:** 152.

Baumgartner, H. and J. B. Steenkamp (2001). "Response Styles in Marketing Research: A Cross - National Investigation." Journal of Marketing Research **37**(May): 143 - 156.

Bay, S. D. and M. Schwabacher (2003). "Mining Distance Based Outliers in Near Line Time with Randomization and a Simple Pruning Rule." KDD'03.

Becken, S. and J. Gnoth (2004). "Tourist consumption systems among overseas visitors: reporting on American, German, and Australian visitors to New Zealand." Tourism Management **25**(3): 375 - 385.

Beckman, R. J. and R. D. Cook (1983). "Outlier...s." Technometrics **25**(May): 119 - 49.

Belk, R. (1986). Generational Differences in the Meaning of Things, Products and Activities. Advertising and Consumer Psychology. K. Sentis and J. C. Olson. New York, Praeger**:** 199 - 213.

Belk, R. W. (1974). "Exploratory Assessment of Situational Effects in Buyer Behavior." Journal of Marketing Research **11**(2): 156-163.

Belk, R. W. (1975). "Situational Variables and Consumer Behavior." Journal of Consumer Research **2**(3): 157 - 165.

Bendre, S. M. and B. K. Kale (1985). "Masking Effect on Tests for Outliers in Exponential Models." Journal of the American Statistical Association **80**(392): 1020-1025.

Bendre, S. M. and B. K. Kale (1987). "Masking Effect on Tests for Outliers in Normal Samples." Biometrika **74**(4): 891-896.

Bernoulli, D. (1777). The Most Probable Choice Between Several Discrepant Observations and the Formation Therefrom of the Most Likely Induction. Biometrika 48. C. G. Allen**:** 3 - 13.

Berry, M. J. A. and G. Linoff (1997). Data Mining Techniques: for marketing, sales and customer support. New York, Wiley.

Bhaskar, R. (1975). A Realist Theory of Science. Brighton, Harvester.

Bhaskar, R. (1986). Scientific Realism and Human Emancipation. London, Verso.

Bhaskar, R. (1998). General Introduction. Critical Realism: Essential Readings. M. Archer. London, Routledge.

Bidwell, C. (1979). "A Test of Market Efficiency: SUE/PE." Journal of Portfolio Management **5**(4): 53 - 58.

Bieger, T. and C. Laesser (2002). "Market Segmentation by Motivation - The Case of Switzerland." Journal of Travel Research **41**: 68 - 76.

Biggs, D., b. de Ville and E. Suen (1991). "A method of choosing multiway partitions for classification and decision trees." Journal of Applied Statistics **18**(1): 49 - 62.

Bird, R., B. McElwee and J. McKinnon (2000). "A Global Perspective of Analysts' Earning Forecasts." Journal of Investing **9**(4): 76 - 82.

Bollen, K. A. (1990). "A Comment on Model Evaluation and Modification." Multivariate Behavioral Research **25**(2): 181-185.

Bolton, R. J. and D. J. Hand (2002). "Statistical fraud detection: A review (with discussion)." Statistical Science **17**(3): 235 - 255.

Bonoma, T. V. and B. P. Shapiro (1984). "Evaluating Market Segmentation Approaches." Industrial Marketing Management **13**(4): 257 - 268.

Borokowski, S. C. and M. J. Welsh (2000). "Ethical practice in the accounting publishing process: Contrasting opinions of authors and editors." Journal of Business Ethics **25**: 15 - 31.

Bortz, J. and N. Döring (2002). Forschungsmethoden und Evaluation. Berlin, Springer.

Box, G. E. P. and G. C. Tiao (1968). "A Bayesian Approach to Some Outlier Problems." Biometrika **55**: 119 - 129.

Bradu, D. and D. M. Hawkins (1985). "Location of Outliers in Regression - Use of Some EM Algorithms." South African Statistical Journal **19**(2): 132-132.

Braithwaite, J. (1989). Crime, Shame and Reintegration. New York, Cambridge University Press.

Braithwaite, J. (2000). "Shame and criminal justice." Canadian Journal of Criminology **42**(July): 281 - 299.

Breckeridge, J. N. (1993). Validating Cluster Analysis: Consistent replication and symmetry. Meeting of the Classification Society of North America, University of Pittsburg, PA.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984). Classification and Regression Trees. Belmont, CA, Wadsworth.

Breunig, M. M., H. P. Kriegel, R. T. Ng and J. Sander (2000). LOF: Identifying Density - Based Local Outliers. Proceedings of the SIGMOD, Dallas, TX.

Brodie, R. J. and P. J. Danaher (2000). "Building models for marketing decisions: Improving empirical procedures." International Journal of Research in Marketing **17**(2): 135 - 139.

Broen, W. E. and R. D. Wirt (1958). "Varieties of Response Sets." Journal of Consulting Psychology **22**(3): 237 - 240.

Brown, B. M. (1975). "A Short-Cut Test for Outliers Using Residuals." Biometrika **62**(3): 623 - 629.

Brown, H. (1977). Perception, Theory and Commitment. Chicago, University of Chicago Press.

Brown, M. B. (1974). "Identification of Sources of Significance in 2-Way Contingency-Tables." Journal of the Royal Statistical Society Series C-Applied Statistics **23**(3): 405-413.

Bruner, J. S. and M. C. Potter (1964). "Inference in visual recognition." Science(144): 424 - 425.

Budescu, D. V., I. Erev and T. S. Wallsten (1997). "On the Importance of Random Error in the Study of Probability Judgement. Part I: New Theoretical Developments." Journal of Behavioral Decision Making **10**(3): 157 - 171.

Burgess, R. L. and R. L. Akers (1966). "A Differential Association - Reinforcement Theory of Criminal Behavior." Social Problems **14**(2): 128 - 147.

Burkhardt, H. and B. Smith (1991). Handbook of Metaphysics and Ontology. Munich, Philosophia Verlag.

Burton, D. (2005). "Marketing theory matters." British Journal of Management **16**(1): 5-18.

Butler, R. (1983). "Outlier Discordancy Tests in the Normal Linear-Model." Journal of the Royal Statistical Society Series B-Methodological **45**(1): 120-132.

Calder, B., L. W. Phillips and A. M. Tybout (1981). "Designing Research for Applications." Journal of Consumer Research **8**(September): 197 - 207.

Calinski, T. and J. Harabasz (1974). "A Dendrite Method for Cluster Analysis." Communications in Statistics **3**: 1-27.

Cannon, J. P. and W. D. Perreault (1999). "Buyer - Seller relationships in Business Markets." Journal of Marketing Research **36**(4): 439 - 460.

Cao, L. J., H. P. Lee and W. K. Chong (2003). "Modified support vector novelty detector using training data with outliers." Pattern Recognition Letters **24**(14): 2479 - 2487.

Carnap, R. (1936). "Testability and Meaning." Philosophy of Science **3**: 419 - 471.

Carnap, R. (1937). "Testability and Meaning." Philosophy of Science **4**: 1 - 40.

Carnap, R. (1946). "Theory and Prediction in Science." Science(104): 520-521.

Carnap, R. (1953). Testability and Meaning. Readings in the Philosophy of Science. H. Feigel and M. Brodbeck. New York, Appleton - Century - Crofts**:** 47 - 92.

Carnap, R. (1956). The Methodological Character of Theoretical Concepts. The Foundations of Science and the Concepts of Psychology and Psychoanalysis. H. Feigl and M. Scriven. Minneapolis, University of Minnesota Press. **1:** 38 - 76.

Carol, R. J. (1980). "Robust Methods for Factorial Experiments with Outliers." Applied Statistics **29**(3): 246 - 251.

Caroll, R. J. (1982). "Two Examples of Transformations When There Are Possible Outliers." Applied Statistics **31**: 149 - 152.

Carson, D., I. Waller and N. Scott (2002). Drive Tourism: Up the Wall and Around the Bend. Common Ground Publishing, Melbourne.

Carter, C. K. and R. Kohn (1996). Robust Bayesian nonparametirc regression. Statistical Theory and Computational Aspects of Smoothing. W. Hardle and M. G. Schimek. Heidelberg, Physica - Verlag.

Castillo, E., A. S. Hadi, N. Balakrishnan and J. M. Srabia (2005). Extreme Value and Related Models with Applications in Engineering and Science. Hoboken, NJ, Wiley.

Causey, R. L. (1979). Theory and Observation. Current Research in Philosophy. P. D. Asquith and H. E. Kyburg. East Lansing, MI, Philosophy of Science Association**:** 187 - 206.

Chamberlin, T. C. (1890). "The method of multiple working hypotheses." reprinted in Science 1965 **148**: 754 - 759.

Chambers, R. L. and C. R. Heathcote (1981). "On the Estimation of Slope and the Identification of Outliers in Linear-Regression." Biometrika **68**(1): 21-33.

Chambliss, W. J. (1969). Crime and the Legal process. Reading, MA, Addison - Wesely.

Chambliss, W. J. and R. Seidman (1971). Law, Order and Power. Reading, MA, Addison - Wesely.

Chan, W. S. (1995). "Understanding the effect of time series outliers on sample autocorrelation." Test **4**(179 - 186).

Chan, W. S. and S. H. Cheung (1994). "On Robust Estimation of Threshold Autoregressions." Journal of Forecasting **13**(1): 37-49.

Chang, I., G. C. Tiao and C. Chen (1988). "Estimation of Time-Series Parameters in the Presence of Outliers." Technometrics **30**(2): 193-204.

Chase, C. W. J. (2001). "Understanding and using the power of outliers in statistical modelling." Journal of Business Forecasting Methods & Systems **19**(4): 2.

Chatterjee, S. and A. S. Hadi (1988). Sensitivity analysis in linear regression, Wiley series in probability and mathematical statistics.

Chaturvedi, A., J. D. Carroll, P. E. Green and J. A. Rotondo (1997). "A feature-based approach to market segmentation via overlapping K-centroids clustering." Journal of Marketing Research **34**(3): 370-377.

Chen, C., S. Y. Lee and H. W. Stevenson (1995). "Response Style and Cross-Cultural Comparisons of Rating-Scales among East-Asian and North-American Students." Psychological Science **6**(3): 170-175.

Chen, C. and L. M. Liu (1993). "Joint Estimation of Model Parameters and Outlier Effects in Time-Series." Journal of the American Statistical Association **88**(421): 284-297.

Chernick, M. R., D. J. Downing and D. H. Pike (1982). "Detecting Outliers in Time-Series Data." Journal of the American Statistical Association **77**(380): 743-747.

Chow, W. W. (2004). "An outlier robust hierarchical Bayes model for forecasting: The case of Hong Kong." Journal of Forecasting **23**(2): 99-114.

Chubachi, S. (1984). "Preliminary Result of Ozone Observation at Syowa Station from February 1982 to January 1983." Memoirs of the National Institute for Polar Research Special Issue **34**(1984): 13 - 19.

Clancy, K. J. and R. S. Shulman (1994). Marketing Myths that are Killing Business. New York, McGraw - Hill.

Clark, T. (1989). "Managing Outliers: Qualitative Issues in the Handling of Extreme Observations in Marketing Research." Marketing Research **26**(June): 31 - 48.

Clarke, N. (2004). "Free independent travellers? British working holiday makers in Australia." Transactions of the Institute of British Geographers **29**(4): 499 - 509.

Cleveland, W. S. (1979). "Robust Locally Weighted Regression and Smoothing Scatterplots." Journal of the American Statistical Association **74**(368): 829-836.

Cleveland, W. S. (1984). "Graphical Methods for data presentation: Full scale breaks, dot charts, and multibased logging." The American Statistician **38**(4): 270 - 280.

Cleveland, W. S. (1993). Visualizing data. Hobart, Summit.

Cleveland, W. S. and M. E. McGill (1988). Dynamic graphics for statistics. Belmont, CA, Wadsworth.

Cloward, R. and L. E. Ohlin (1960). Delinquency and Opportunity: A theory of Juvenile Gangs. New York, Free Press.

Cohen, J. (1968). "Multiple regression as a general data-analytic system." Psychological Bulleting **70**: 426 - 443.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillsdale, NJ, Lawrence Earlbaum Associates.

Cohen, J. (1992). "A Power Primer." Psychological Bulletin **112**(1): 155-159.

Cohen, J. (1994). "The earth is round (p<.05)." American Psychologist **49**: 997 - 1003.

Cohen, L. E. (1955). Delinquent Boys: The Culture of the Gang. New York, Free Press.

Collett, D. and T. Lewis (1976). "Subjective Nature of Outlier Rejection Procedures." Journal of the Royal Statistical Society Series C-Applied Statistics **25**(3): 228-237.

Collier, A. (1994). An Introduction to Roy Bhaskar's Philosophy. London, Verso.

Comte, A. (1842). Cours de philosophie positive. Paris, Borrani et Dros, Libraires.

Cook, T. D. and D. T. Campbell (1979). Quasi-Experimentation - Design & Analysis Issues for Field Settings. Chicago, Rand Mc Nally College Publishing Company.

Cook, V. J. and W. A. Mindak (1984). "A search for constants: The "Heavy User" Revisited!" Journal of Consumer Marketing **1**(4): 79 - 81.

Couch, A. and K. Keniston (1960). "Yeasayers and Naysayers - Agreeing Response Set as a Personality Variable." Journal of Abnormal and Social Psychology **60**(2): 151-174.

Court, B. and R. A. Luption (1997). "Customer Portfolio Development: Modelling Destination Adopters, Inactives, and Rejecters." Journal of Travel Research **36**(1): 35 - 43.

Cowles, M. (2001). Statistics in Psychology - an historical perspective. Mahwah, NJ, Lawrence Erlbaum Associates Inc.

Cronbach, L. J. (1946). "Response Set and Test Validity." Educational and Psychological Measurement **6**(Winter): 475 - 494.

Cronbach, L. J. (1950). "Further Evidence on Response Sets and Test Design." Educational and Psychological Measurement **10**: 3 - 31.

Cronbach, L. J. (1957). "The two Disciplines of Scientific Psychology." American Psychologist **12**: 671 - 684.

Davidson, J. O. C. and D. Layder (1994). Methods, Sex and Madness. London, Routledge.

Denby, L. and R. D. Martin (1979). "Robust Estimation of the 1st-Order Autoregressive Parameter." Journal of the American Statistical Association **74**(365): 140-146.

Desphande, R. (1983). "Paradigms Lost: On Theory and Method in Research in Marketing." Journal of Marketing **47**(4): 101 - 110.

Dijk, v., P. H. Franses and A. Lucas (1999). "Testing for ARCH in the presence of additive outliers." Journal of Applied Econometrics **14**: 539 - 562.

Dixon, W. J. (1962). Rejection of Observations. Contributions to Order Statistics. A. E. Sahron and B. G. Greenberg. New York, John Wiley & Sons Inc.

Dodd, T. and V. Bigotte (1997). "Perceptual Differences among Visitor Groups to Wineries." Journal of Travel Research **35**(3): 46 - 51.

Dolnicar, S. (2004). "Beyond "Commonsense Segmentation": A Systematics of Segmentation Approaches in Tourism." Journal of Travel Research **42**(3): 244 - 250.

Dornbos, R. (1976). Slippage Tests. Amsterdam, Mathematisch Centrum.

Douglas, C. J. and B. Isherwood (1979). The World of Goods. New York, Norton.

Doyle, P. and I. Fenwick (1975). "Pitfalls of AID analysis." Journal of Marketing Research **12**(November): 408 - 413.

Dubois, P.-L., A. Jolibert and H. Mühlbacher (2007). Marketing Management. Houndsmill and New York, Palgrave MacMillan.

Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis. New York, John Wiley&Sons Inc.

Duhem (1953). Physical Theory and Experiment. Readings in the Philosophy of Science. H. Feigl and M. Brodbeck. New York, Appleton - Century - Crofts**:** 235 - 252.

Duhem, P. (1962). The Aim and Structure of Physical Theory. New York, Atheneum.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. The Nature of Insight. R. J. Steinberg and J. E. Davidson. Cambridge, MA, MIT Press.

Dunning, D., D. W. Griffin, J. D. Milojkovic and L. Ross (1990). "The Overconfidence Effect in Social Prediction." Journal of Personality and Social Psychology **58**(4): 568-581.

Dupuis, D. J. and D. C. Hamilton (2000). "Regression residuals and test statistics: assessing naive outlier deletion." Canadian Journal of Statistics-Revue Canadienne De Statistique **28**(2): 259-275.

Dwyer, L. and P. Forsyth (1998). "Economic significance of cruise tourism." Annals of Tourism Research **25**(2): 393 - 415.

Easterby-Smith, M. and R. Thorpe (2002). Management Research - An Introduction. London, Sage.

Easton, G. (1998). Case Research as Methodology for Industrial Networks: A Realist Apologia. Network Dynamics in International Marketing. P. Naudé and P. W. Turnbull. Oxford, Elsevier Science Ltd.

Easton, G. (2002). "Marketing - A critical realist approach." Journal of Business Research **55**(2): 103 - 109.

Ebbinghaus, B. (2005). Not Just Outliers: Making Sense of Exceptionalism in Cross-national Comparisons. ECPR Conference, Budapest.

Emerson, J. D. and D. C. Hoaglin (1983). Analysis of Two Way Tables by Medians. Understanding Robust and Exploratory Data Analysis. D. C. Hoaglin, F. Mosteller and J. W. Tukey. New York, John Wiley & Sons, Inc.

Engel, J. F., R. D. Blackwell and P. W. Miniard (1994). Consumer Behavior. Fort Worth, Dryden Press.

Erikson, K. T. (1962). "Notes on the sociology of deviance." Social Problems **9**: 307 - 314.

Ermann, M. D. and R. J. Lundman (2002). Corporate and Governmental Deviance: Problems of Organizational Behavior in Contemporary Society. New York, Oxford University Press.

Eskin, E., A. Arnold, M. Prera, L. Portnoy and S. Stolfo (2002). "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data." Datamining for Security Applications.

Everitt, B. S., S. Landau and M. Leese (2001). Cluster Analysis. London, Arnold.

Farman, J. C., B. G. Gardiner and J. D. Shanklin (1985). "Large Losses of Total Ozone Reveal Seasonal ClOx/NOx Interaction." Nature **315**(1985): 207 - 210.

Farris, H. and R. Revlin (1989). "The discovery process: A conterfactual strategy." Social Studies of Science **19**(497 - 513).

Ferguson, T. S. (1961a). "On the Rejection of Outliers." Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability **1**(253 - 287).

Ferguson, T. S. (1961b). "Rules for Rejection of Outliers." Review de l'Institut International de Statistique **29**: 29 - 43.

Ferro, C. A. T. and J. Segers (2003). "Inference for clusters of extreme values." Journal of the Royal Statistical Society: B **65**(2): 545 - 556.

Fesenmaier, D. R. and B. Johnson (1989). "Involvement-based segmentation: Implications for travel marketing in Texas." Tourism Management **10**(4): 293 - 300.

Feyerabend, P. K. (1978). Against Method: Outline of an Anarchistic Theory of Knowledge. London, Verso.

Field, A. (2000). Discovering statistics using SPSS. London, Sage.

Field, A. M. (1999). "The College Student Market Segment: A Comparative Study of Travel Behaviors of International and Domestic Students at a Southeastern University." Journal of Travel Research **37**(4): 375 - 381.

Fieller, N. R. J. (1976). Some Problems Related to the Rejection of Outlying Observations. unpublished PhD Thesis. Sheffield, University of Sheffield.

Finkleman, B. (1974). "Ads Should Reinforce Current Users, Not Necessarily Convert Nonusers of Products." Marketing News(January 15, 1974): p.1.

Finney, D. J. (2006). "Calibration guidelines challenge outlier practices." American Statistician **60**(4): 309-314.

Fishbein, M. and I. Ajzen (1975). Belief, attitude, intention and behavior: An introduction to theory and research. Reading, MA, Addison-Wesley.

Fisher, R. A. (1925a). "Applications of "Student's" distribution." Metron **5**: 90 - 104.

Fisher, R. A. (1925b). "Theory of statistical estimation." Proceedings of the Campridge Philosophical Society **22**: 700 - 725.

Fleetwood, S. (2001). "Causal Laws, Functional Relations and Tendencies." Review of Political Economy **13**(2): 201 - 220.

Fodness, D. and B. Murray (1998). "A Typology of Tourist Information Search Strategies." Journal of Travel Research **37**(2): 108 - 119.

Fodness, D. and B. Murray (1999). "A Model of Tourist Information Search Behavior." Journal of Travel Research **37**: 220 - 230.

Forminca, S. and M. Uysal (1998). "Market Segmentation of an International Cultural-Historical Event in Italy." Journal of Travel Research **36**(4): 16 - 24.

Fox, A. J. (1972). "Outliers in Time Series." Journal of the Royal Statistical Society Series B-Statistical Methodology **34**(3): 350-363.

Frank, R. E., W. F. Massy and Y. Wind (1972). Market Segmentation. Englewood Cliffs, NJ, Prentice Hall.

Franke, N., H. Reisinger and D. Hoppe (2007). Remaining within - cluster Variance: A Meta - Analysis of The "Dark Side" of Cluster Analysis. Working Paper - Vienna University of Economics and Business Administration. Vienna.

Franke, N. and E. Von Hippel (2003). "Satisfying Heterogeneous User Needs via Innovation Toolkits:

The Case of Apache Security Software." Research Policy **32**: 1199 - 1215.

Freeman, P. R. (1980). On the Number of Outliers in Data From a Linear Model. Bayesian Statistics. J. M. Bernardo, M. H. DeGrood, D. V. Lindley and A. F. M. Smith. Valencia, Valencia University Press**:** 349 - 365.

Fuchs, C. and R. Kenett (1980). "A Test for Detecting Outlying Cells in the Multinomial Distribution and 2-Way Contingency-Tables." Journal of the American Statistical Association **75**(370): 395-398.

Galambos, J. and N. Macri (2002). "Classical extreme value model and prediction of extreme winds." Journal of Structural Engineering **125**: 792 - 794.

Galpin, J. S. and D. M. Hawkins (1981). "Rejection of a Single Outlier in 2-Way or 3-Way Layouts." Technometrics **23**(1): 65-70.

Gans, J. S. and G. B. Shepard (1994). "How are the mighty fallen: Rejected classic articles by leading economists." Journal of Economic Perspectives **8**: 165 - 179.

Gigerenzer, G. (1991). "From tools to theories: A heuristic of discovery in cognitive psychology." Psychological Review **98**: 254 - 267.

Gigerenzer, G. and U. Hoffrage (1995). "How to improve Bayesian reasoning without instruction: Frequency formats." Psychological Review **102**: 684 - 702.

Gigerenzer, G., S. Krauss and O. Vitouch (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing but Were Afraid to Ask. The SAGE handbook of quantitative methodology for the social. D. Kaplan. Thousand Oaks, Sage.

Giltinan, D. M., R. J. Carroll and D. Ruppert (1986). "Some New Estimation Methods for Weighted Regression When There Are Possible Outliers." Technometrics **28**(3): 219-230.

Gladwell, M. (2000). The tipping point: How little things can make a big difference. Boston, Little, Brown and Company.

Glaister, E. M. and P. Glaister (2004). "Line fitting with Outliers." Teaching Statistics **26**(3): 78 - 80.

Glaser, D. (1956). "Criminality theories and behavioral images." American Journal of Sociology **61**: 423 - 449.

Glass, G. V., B. McGaw and M. L. Smith (1981). Meta Analysis in Social Research. Beverly Hills, CA, Sage.

Glynn, P. W. and W. Whitt (1995). "Heavy traffic extreme value limits for queues." Operations Research Letters **18**(3): 107 - 111.

Gnanades.R and Kettenri.Jr (1972). "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data." Biometrics **28**(1): 81-&.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. New York, Wiley.

Gnoth, J. (1999). Tourism expectation formation: The case of Camper Van Tourists in New Zealand. Consumer Behavior in Travel and Tourism. A. Pizam and Y. Mansfeld. New York, London, Oxford, The Haworth Hospitality Press.

Gokhale, D. V. and S. Kullback (1978). The Information in Contingency Tables. New York, Marcel Dekker.

Goldsmith, R. E. (2000). "Characteristics of the heavy user of fashionable clothing." Journal of Marketing Theory and Practice **8**(4): 21 - 28.

Goldsmith, R. E. and S. W. Litvin (1999). "Heavy Users of Travel Agents: A Segmentation Analysis of Vacation Travelers." Journal of Travel Research **38**(2): 127.

Gómez-Corral, A. (2001). "On extreme values of orbit lengths in M/G/1 queues with constant retrial rate." OR Spektrum. Quantitative Approaches in Management **23**(3): 395 - 409.

Goodman, L. A. (1974). "Exploratory Latent Structure-Analysis Using Both Identifiable and Unidentifiable Models." Biometrika **61**(2): 215-231.

Gorman, M. E. and M. Gorman (1984). "A comparison of disconfirmatory, confirmatory and control strategies on Wason's 2-4-6 task." The Quarterly Journal of Experimental Psychology **36A**: 629 - 648.

Goulding, C. (1999). "Consumer research, interpretive paradigms and methodological ambiguities." European Journal of Marketing **33**(9/10): 859 - 873.

Green, R. F. (1976). "Outlier-Prone and Outlier-Resistant Distributions." Journal of the American Statistical Association **71**(354): 502-505.

Greenberg, D. F. (1981). Crime and Capitalism: Readings in Marxist Criminology. Palo Alto, Mayfield.

Greenleaf, E. A. (1992). "Measuring Extreme Response Style." Public Opinion Quarterly **56**(3): 328-351.

Greenwald, A. G. (1975). "Consequences of prejudice against the null hypothesis." Psychological Bulletin **82**: 1 - 20.

Greenwald, A. G., M. R. Leippe, A. R. Pratkanis and M. H. Baumgardner (1986). "Under What Conditions Does Theory Obstruct Research Progress." Psychological Review **93**(2): 216 - 229.

Greenwald, A. G. and A. R. Pratkanis (1988). "On the use of theory and the usefulness of theory." Psychological Review **93**(2): 216 - 229.

Greenwald, A. G. and D. L. Ronis (1981). "On the Conceptual Disconfirmation of Theories." Personality and Social Psychology Bulletin **7**(1): 131-137.

Groves, R. M. (1991). Measurement errors across the disciplines. Measurement errors in surveys. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz and S. Sudman. New York, John Wiley & Sons.

Grubbs, F. E. (1969). "Procedures for Detecting Outlying Observations in Samples." Technometrics **11**(1): 1-&.

Gummesson, E. (2003). "All research is interpretive." Journal of Business & Industrial Marketing **18**(6/7): 482 - 492.

Guttman, I. (1973a). "Care and handling of univariate or multivariate outliers in detecting spuriosity - a Bayesian approach." Technometrics **15**: 723 - 738.

Guttman, I. (1973b). "Care and Handling of Univariate or Multivariate Outliners in Detecting Spuriosity - Bayesian Approach." Technometrics **15**(4): 723-738.

Haberman, S. J. (1973). "Analysis of Residuals in Cross-Classified Tables." Biometrics **29**(1): 205-220.

Hadi, A. S. and J. S. Simonoff (1993). "Procedures of identification of multiple outliers in linear models." Journal of the American Statistical Association **88**: 1264 - 1272.

Hahn, C. H. (2002). Segmentspezifische Kundenzufriedenheitsanalyse - Neue Ansätze zur Segmentierung von Märkten. Wiesbaden, Deutscher Universitäts - Verlag.

Hamilton, D. L. (1968). "Personality Attributes Associated with Extreme Response Style." Psychological Bulletin **69**(3): 192-&.

Handel, W. (1982). Ethnomethodology: How people make sense. Englewood Cliffs, NJ, Prentice Hall.

Harkins, S., H. He, G. Williams and R. A. Baster (2002). "Outlier Detection Using Replicator Neural Networks." DaWaK'02: 170 - 180.

Harlow, L. L. (1997). Significance Testing Introduction and Overview. What if there were No Significance Tests? L. L. Harlow, S. A. Mulaik and J. H. Steiger. Mahwah, NJ, Lawrence Erlbaum Associates.

Harter, H. L. (1977). A Survey of the Literature on the Size Effect on Material Strength. Wright Patterson Air Force Base Ohio, Air Force Flight Dynamics Laboratory/FBRD, Air force Systems Command, Unites States Air Force.

Harter, H. L. (1978). "A bibliography of extreme value theory." International Statistical Review **46**(279 - 306).

Hawkins, D. M. (1980). Identification of Outliers. London, Chapman and Hall.

He, X., X. Xu and S. Deng (2003). "Discovering cluster based local outliers." Pattern Recognition Letters **24**: 1641 - 150.

He, X., X. Xu, J. Z. Huang and S. Deng (2004). "A Frequent Pattern Discovery Based Method of Outlier Detection." WAIM'04.

He, Z., X. Xu and S. Deng (2005). An Optimization Model for Outlier Detection in Categorical Data. ICIC 2005, Hefei, Advances in intelligent computing.

Hempel, C. G. (1965). Aspects of Scientific Explanation. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. New York, The Free Press**:** 331 - 496.

High, R. (2000). Dealing with 'Outliers': How to Maintain Your Data's Integrity. Computing News, University of Oregon. Eugene, OR.

Hirschi, T. and M. Gottfredson (1994). The Generality of Deviance. New Brunswick, NJ, Transaction.

Hirschi, T. and M. Gottfredson (2000). "In Defense of self - control." Theoretical Criminology **4**: 55 - 69.

Hjoerland, B. (2004). "Empiricism, rationalism and positivism in library and informations science." Journal of Documentation **61**(1): 130 - 155.

Hoaglin, D. C., B. Iglewicz and J. W. Tukey (1986). "Performance of Some Resistant Rules for Outlier Labelling." Journal of the American Statistical Association **81**(396): 991-999.

Hoch, S. and Y.-W. Ha (1986). "Consumer Learning: Advertising and the Ambiguity of Product Experience." Journal of Consumer Research **13**(2): 221 - 234.

Hofstede, G. H. (1984). Culture's Consequences: International Differences in Work - Related Values, Sage Publications Inc.

Hogg, M. K. and P. C. N. Mitchell (1997). "Constellations, Configurations and Consumption: Exploring Patterns of Consumer Behaviour Amongst UK Shoppers." Advances in Consumer Research **24**: 551 - 559.

Holbrook, M. B. and R. Batra (1991). "Consumer Responses to Advertising: The Effect of Ad Content, Emotions, and Attitude toward the Ad on Viewing time." Journal of Consumer Research **17**(4): 440 - 454.

Horneman, L., R. W. Carter, S. Wei and H. Ruys (2002). "Profiling the Senior Traveler: An Australian Perspective." Journal of Travel Research **41**(1): 23 - 37.

Howard, J. A. and J. N. Sheth (1969). The Theory of Buyer Behavior. New York, Wiley.

Hsu, C. H. C. and E.-J. Lee (2002). "Segmentation of Senior Motorcoach Travellers." Journal of Travel Research **40**(4): 364 - 374.

Hu, T. and S. Y. Sung (2003). "Detecting pattern-based outliers." Pattern Recognition Letters **24**(16): 3059 - 3068.

Huba, G. J. (2003). CHAID definition, the measurement group. **2004**.

Huba, G. J. and L. L. Harlow (1987). "Robust Structural Equation Models - Implications for Developmental-Psychology." Child Development **58**(1): 147-166.

Hubbard, R. and G. Armstrong (1992). "Are null results becoming an endangered species in marketing?" Marketing Letters **3**: 127 - 136.

Hubbard, R. and G. Armstrong (1994). "Replications and Extensions in Marketing-Rarely Published, But Quite Contrary." International Journal of Research in Marketing **11**: 233 - 248.

Hubbard, R. and J. S. Armstrong (1997). "Publication bias against null results." Psychological Reports **80**(1): 337-338.

Hubbard, R. and R. M. Lindsay (2002). "How the emphasis on 'original' empirical marketing research impedes knowledge development." Marketing Theory **2**(4): 381 - 402.

Huber, P. J. (2004). Robust Statistics. Hoboken, NJ, John Wiley & Sons.

Huck, S. W. (2000). Reading statistics and research. New York, Longman.

Hui, C. H. and H. C. Triandis (1985). "The Instability of Response Sets." Public Opinion Quarterly **49**(2): 253-260.

Hume, D. (1777). Enquiries Concerning Human Understanding and Concerning the Principles of Morals. Reprinted from 1777 edition. L. A. Selby-Bigge. Oxford, Clarendon Press. **Sect. XII PartII**.

Hume, D. (1955). An Inquiry Concerning Human Understanding (1748). Indianapolis, Bobbs - Merrill.

Hunt, J. (2003). Controversy in Marketing Theory. New York, M.E. Sharpe.

Hunt, S. (1976). "General theories and the fundamental explanada of marketing." Journal of Marketing **47**: 9 - 17.

Hunt, S. (1990). "Truth in Marketing Theory and Research." Journal of Marketing **54**(3): 1 - 16.

Hunt, S. (1991a). Modern Marketing Theory: Critical Issues in the Philosophy of Marketing Science. Cincinnati, OH, South Western.

Hunt, S. (1991b). "Positivism and Paradigm dominance in Consumer Research: Toward Critical Pluralism and Rapprochement." Journal of Consumer Research **18**: 32 - 44.

Hunt, S. (1994). "On Rethinking Marketing: Our Discipline, Our Practice, Our Methods." European Journal of Marketing **28**(3): 13 - 25.

Hunter, J. E., F. L. Schmidt and G. B. Jackson (1982). Meta Analysis: Cumulating Research Findings Across Studies. Beverly Hills, CA, Sage.

Iglewicz, B. and D. C. Hoaglin (1993). How to detect and handle outliers. Milwaukee, WI, ASQC Quality Press.

Israeli, A. A. (2002). "A Preliminary Investigation of Importance of Site Accessibility Factors for Disabled Tourists." Journal of Travel Research **39**(1): 45 - 51.

Jackson, D. N. (1967). Acquiescence Response Styles: Problems of Identification and Control. Response Set in Personality Assessment. I. A. Berg. Chicago, Aldine Publishing Company**:** 71 - 114.

Jasso, G. (1986). "Is It Outlier Deletion or Is It Sample Truncation - Notes on Science and Sexuality - Reply." American Sociological Review **51**(5): 738-742.

Jennings, D. E. (1986). "Outliers and Residual Distributions in Logistic-Regression." Journal of the American Statistical Association **81**(396): 987-990.

Jiang, M. F., S. S. Tseng and C. M. Su (2001). "Two -phase Clustering Process for Outlier Detection." Pattern Recognition Letters **22**(6-7): 387 - 412.

Jin, W., A. K. H. Tung and J. Han (2001). "Mining top-n local outliers in large databases." KDD'01.

Johansson, I. (1975). A Critique of Karl Poppper's Methodology. Stockholm, Akademiförlaget.

John, J. A. (1978). "Outliers in Factorial Experiments." Applied Statistics **27**: 111 - 119.

Johnson, T., I. Kwok and R. T. Ng (1998). "Fast computation of 2-dimensional depth contours." KDD'98.

Joshi, P. C. (1972). "Efficient Estimation of Mean of an Exponential Distribution When an Outlier Is Present." Technometrics **14**(1): 137-&.

Judd, C. M. and D. C. McClelland (1989). Data Analysis: A model comparison approach. San Diego, CA, Harcourt Brace Jovanovich.

Judd, C. M., G. H. McClelland and S. E. Culhane (1995). "Data-Analysis - Continuing Issues in the Everyday Analysis of Psychological Data." Annual Review of Psychology **46**: 433-465.

Julander, C. R. and M. Söderlund (2002). What about the cases that did not conform to the accepted hypotheses - are they noise or useful information. EMAC, Braga, Portugal.

Kabe, d. G. (1970). "Testing Outliers from and Exponential Population." Metrika **15**: 15 - 18.

Kahn, J. R. and J. R. Udry (1986). "Marital Coital Frequency - Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions." American Sociological Review **51**(5): 734-737.

Kara, A. and E. Kaynak (1997). "Markets of a single customer: exploiting conceptual development in market segmentation." European Journal of Marketing **31**(11/12): 873 - 895.

Karr, A. F. (1976). "Two extreme value processes arising in hydrology." Journal of Applied Probability **13**(1): 190 - 194.

Kashyap, R. and D. C. Bojanic (2000). "A Structural Analysis of Value, Quality and Price Perceptions of Business and Leisure Travellers." Journal of Travel Research **39**(1): 45 - 51.

Kass, G. V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data." Journal of the Royal Statistical Society: Series C (Applied Statistics) **29**(2): 119 - 128.

Kastenholz, E., D. Davis and G. Paul (1999). "Segmenting tourism in rural areas: The Case of North and Central Portugal." Journal of Travel Research **37**(4): 353 - 363.

Katz, J. (1988). Seductions of Crime: Moral and Sensual Attractions in Doing Evil. New York, Basic Books.

Kenny, A. (1986). Rationalism, Empiricism, and Idealism. Oxford, Oxford University Press.

Kerr, S. J., J. Tolliver and D. Petree (1977). "Manuscript characteristics which influence acceptance for management and social science journals." Academy of Management Journal **20**: 132 - 141.

Ketchen, D. J. and C. L. Shook (1996). "The application of cluster analysis in strategic management research: An analysis and critique." Strategic Management Journal **17**(6): 441-458.

Kimber, A. C. (1983). "Trimming in Gamma Samples." Applied Statistics **32**: 7 - 14.

King, E. P. (1953). "On Some Procedures for the Rejection of Suspected Data." Journal of the American Statistical Association **48**(263): 531-533.

Kirk, R. E. (1996). "Practical Significance: A concept whose time has come." Educational and Psychological Measurement **56**(746 - 759).

Klecka, W. R. (1980). Discriminant Analysis. Sage University Paper series on Quantitative Applications in the Social Sciences. Beverly Hills and London, Sage Publications. **07-44**.

Klemm, M. S. (2002). "Tourism and Ethnic Minorities in Bradford: The Invisible Segment." Journal of Travel Research **41**(1): 85 - 91.

Kline, R. (2005). Principles and practices of structural equation modelling. New York, The Guilford Press.

Knorr, E. M. and R. T. Ng (1997). "A unified notion of outliers: Properties and computation." KDD'97: 219 - 222.

Knorr, E. M. and R. T. Ng (1998). "Algorithms for mining distance based outliers in large datasets." VLDB'98.

Knorr, E. M. and R. T. Ng (1999). Finding intentional knowledge of distance - based outliers. Proceedings of the 25th VLDB Conference.

Knorr, E. M., R. T. Ng and V. Tucakov (2000). "Distance based outliers: Algorithms and Applications." The VLDP Journal **8**: 237 - 253.

Knowles, E. S. and K. T. Nathan (1997). "Acquiescent responding in self-reports: Cognitive style or social concern?" Journal of Research in Personality **31**(2): 293-301.

Kohonen, T. (2001). Self-Organizing Maps. Berlin, Springer.

Kotler, P. (1986). "The Prosumer Movement." Advances in Consumer Research **13**(1): 510 - 514.

Kotler, P. and K. L. Keller (2006). Marketing Management. London, Prentice Hall.

Kotze, T. J. V. and D. M. Hawkins (1984). "The Identification of Outliers in 2-Way Contingency-Tables Using 2x2 Subtables." Applied Statistics-Journal of the Royal Statistical Society Series C **33**(2): 215-223.

Krausz, M. and J. W. Meiland (1982). Relativism: Cognitive and Moral. Notre Dame, IL, University of Notre Dame Pess.

Kuhn, T. S. (1970). The Structure of Scientific Revolutions. Chicago, University of Chicago Press.

Lakatos, I. (1974), “Falsification and the Methodology of Scientific Research Programs”, Criticism and the Growth of Knowledge, I. Lakatos and A. Musgrave(eds), Cambridge: Cambridge University Press (91 – 195)

Lance, C. E., A. M. Stewart and T. R. Carretta (1996). "On the Treatment of Outliers in cognitive and Psychomotor Test Data." Military Psychology **8**(1): 43 - 58.

Lantané, H. and C. P. Jones (1977). "Standardized Unexpected Earnings - A Progress Report." Journal of Finance **22**(5): 1457 - 1465.

Lantané, H. and C. P. Jones (1979). "Standardized Unexpected Earnings 1971 - 1977." Journal of Finance **34**(3): 717 - 724.

Laudan, L. (1977). Progress and Its Problems. Berkeley, CA, University of California Press.

Laudan, L. (1981). "A Confutation of Convergent Realism." Philosophy of Science **48**(March): 19 - 49.

Lawrence, C. N. (2007). "Of Shirking, Outliers, and Statistical Artefacts: Lame-Duck Legislators and Support for Impeachment." Political Research Quarterly **60**(1): 159 - 164.

Lawson, T. (2001). Economics & Reality. London, Rutledge.

Lazarsfeld, P. F. and N. W. Henry (1968). Latent structure analysis. Boston, Hughton Mifflin.

Leadbetter, M. R., G. Lindgren and H. Rootzén (1983). Extreme and Related Properties of Random Sequences and Processes. New York, Springer.

Leigh, J. H. and T. G. Gabel (1992). "Symbolic Interactionism: Its Effects on Consumer Behavior and Implications for Marketing Strategy." Journal of Services Marketing **6**(3): 5.

Lemieux, J. and L. McAlister (2005). "Handling Missing Values in Marketing Data: A comparison of Techniques." MSI Reports **5**(2): 41 - 59.

Lentz, T. F. (1938). "Acquiescence as a Factor in the Measurement of Personality." Psychological Bulleting **35**(November): 659.

Leone, R. P. and R. L. Schultz (1980). "A study of marketing generalizations." Journal of Marketing **44**(Winter): 10 - 18.

Lin, S. and D. E. Brown (2006). "An outlier-based data association method for linking criminal incidents." Decision Support Systems **41**(3): 604 - 615.

Lindsay, R. M. (1994). "Publication system biases associated with the statistical testing paradigm." Contemporary Accounting Research **11**: 33 - 57.

Little, R. J. A. and D. B. Rubin (1987). Statistical Analysis with Missing Data. New York, Wiley.

Liu, X., G. Cheng and J. X. Wu (2002). "Analyzing Outliers Cautiously." IEEE Transactions on knowledge and data engineering **14**(2): 432 - 437.

Loeb, L. (1981). From Descartes to Hume: Continental Metaphysics and the Development of Modern Philosophy. London, Penguin Books.

Loker - Murphy, L. and P. L. Pearce (1995). "Young Budget Travellers - Backpackers in Australia." Annals of Tourism Research **22**(4): 819-843.

Lord, C. G., L. Ross and M. R. Lepper (1979). "Biased assimilation and attitude polarization: The effect of prior theories on subsequent considered evidence." Journal of Personality & Social Psychology **37**: 2098 - 2109.

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization. New York, Oxford University Press.

Luce, R. D. (1988). "The tools-to-theory hypothesis: Review of G. Gigerenzer and D.J. Murray, "Cognition as intuitive statistics"." Contemporary Psychology **33**: 582 - 583.

MacCoun, R. J. (1998). "Biases in the interpretation and use of research results." Annual Review of Psychology **49**(259 - 287).

MacGuire, W. J. (1983). "A contextualist theory of knowledge: its implications for innovation and reform in psychological research." Advances in Experimental Psychology **16**: 1 - 47.

Maddala, G. S. (1998). Unit roots, cointegration, and structural change. Cambridge, Cambridge University Press.

Mahajan, V., S. Sharma and Y. Wind (1984). "Parameter Estimation in Marketing Models in the Presence of Influential Response Data: Robust Regression and Applications." Journal of Marketing Research **21**(August): 268 - 277.

Mahoney, M. J. (1977). "Publication prejudices: An experimental study of confirmation bias in the peer review system." Cognitive Therapy and Research **1**: 161 - 175.

Malakoff, D. (1999). "Bayes Offers a "New" Way to Make Sense of Numbers." Science **286**(5444): 1460 - 1464.

Manicas, P. T. and P. F. Secord (1983). "Implications for Psychology of the New Philosophy of Science." American Psychologist **38**(4): 399-413.

Marasinghe, M. G. (1985). "A Multistage Procedure for Detecting Several Outliers in Linear Regression." Technometrics **27**(4): 395 - 399.

March, R. and A. G. Woodside (2005). Introduction to the Theory and Investigation of Planned and Realized Consumer Behaviour. Tourism Behavior. Cambridge, MA, CABI publishing.

Marin, G., R. J. Gamba and B. V. Marin (1992). "Extreme Response Style and Acquiescence among Hispanics - the Role of Acculturation and Education." Journal of Cross-Cultural Psychology **23**(4): 498-509.

Markie, P. (2004). Rationalism vs. Empiricism, Stanford Encyclopedia of Philosophy. **2008**.

Marsh, H. W. (1987). The Self-Description Questionnaire 1: Manual and Research Monograph. San Antonio, TX, Psychological Corporation.

Martin, J. D. (1964). "Acquiescence - Measurement and Theory." British Journal of Social and Clinical Psychology **3**(October): 216 - 225.

Martin, M. A. and S. Roberts (2006). "An evaluation of bootstrap methods for outlier detection in least squares regression." Journal of Applied Statistics **33**(7): 703-720.

Maureen, C. (2001). The ozone layer - a philosophy of science perspective. Cambridge, Cambridge University Press.

Maxwell, S. E. (2004). "The persistence of underpowered studies in psychological research: Causes, consequences, and remedies." Psychological Methods **9**(2): 147-163.

McCracken, G. D. (1992). Culture and Consumption. Bloomington, Indiana University Press.

McGee, R. K. (1967). Response Set in Relation to Personality: An Orientation. Response Set in Personality Assessment. I. A. Berg. Chicago, Aldine Publishing Company**:** 1 - 31.

McIntyre, R. M. and R. K. Blashfield (1980). "A Nearest - Centroid Technique for Evaluating the Minimum - Variance Clustering Procedure." Multivariate Behavioral Research **15**: 225 - 238.

McKenzie, C. R. M. (1998). "Taking into account the strength of an alternative hypothesis." Journal of Experimental Psychology **24**: 771 - 792.

McKercher, B. (2001). "A Comparison of Main-Destination Visitors and through Travellers at a Dual Purpose Destination." Journal of Travel Research **39**(4): 433 - 441.

McMillan, R. G. and H. A. David (1971). "Tests for One of 2 Outliers in Normal Samples with Known Variance." Technometrics **13**(1): 75-85.

Mead, G. H. (1934). Mind, Self and Society. Chicago, University of Chicago Press.

Meric, H. J. and J. Hunt (1998). "Ecotourists' Motivational and Demographic Characteristics: A Case of North Carolina Travellers." Journal of Travel Research **36**(4): 57 - 61.

Merton, R. K. (1938). "Social Structure and anomie." American Sociological Review **3**: 672 - 682.

Messick, S. (1967). The Psychology of Acquiescence: An Interpretation or Research Evidence. Response Set in Personality Assessment. I. A. Berg. Chicago, Aldine Publishing Company**:** 115 - 145.

Messick, S. (1968). Response Sets. International Encyclopaedia of the Social Sciences. D. L. Sills. New York, MacMillan. **13:** 492 - 496.

Messick, S. (1991). Psychology and Methodology of Response Styles. Improving Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach. R. E. Snow and D. E. Wiley. Hillsdale, NJ, Lawrence Erlbaum Associates**:** 161 - 200.

Messner, S. F. and R. Rosenfeld (2001). Crime and the American Dream. Belmont, Wadsworth.

Micceri, T. (1989). "The Unicorn, the Normal Curve, and Other Improbable Creatures." Psychological Bulletin **105**(1): 156-166.

Miles, M. and M. Huberman (1994). Qualitative Data. Thousand Oaks, Sage Publications.

Miller, A. G. (1972). The social psychology of the psychological experiments. New York, Free Press.

Miller, J. (1991). "Reaction time analysis with outlier exclusion: Bias varies with sample size." The Quarterly Journal of Experimental Psychology **43**(4): 907 - 912.

Milligan, G. W. and M. C. Cooper (1985). "An Examination of Procedures for Determining the Number of Clusters in a Data Set." Psychometrika **50**(2): 159-179.

Milligan, G. W. and S. C. Hirtle (2003). Clustering and Classification Methods. Handbook of Psychology, Research Methods in Psychology. I. B. Weiner. Hoboken, NJ, John Wiley & Sons, Inc.

Mintzberg, H. (1978). "Patterns in Strategy Formation." Management Science **24**(9): 934-948.

Mir, R. and A. Watson (2001). "Critical realism and constructivism in Strategy Research - Toward a Synthesis." Strategic Management Journal **22**(12).

Mirvalieu, M. (1978). "The Rejection of Outlying Observations in Regression Analysis." Theoretical Probability **23**: 598 - 602.

Mittelstrass, J. (2004). Philosophie und Wissenschaftstheorie. Stuttgart, J.B. Metzler.

Moran, M. A. and R. G. Mcmillan (1973). "Tests for One or 2 Outliers in Normal Samples with Unknown Variance - Correction." Technometrics **15**(3): 637-640.

Morey, L. C., R. K. Blashfield and H. A. Skinner (1983). "A comparison of cluster analysis techniques within a sequential validation framework." Multivariate Behavioral Research **18**: 309 - 329.

Morris, M. B. (1977). An Excursion into Creative Sociology. New York, Columbia University Press.

Morrison, J. E. and J. A. Smith (2001). "Scaling Properties of flood peaks." Extremes **4**(1): 5 - 22.

Morrison, P. D., J. H. Roberts and E. Von Hippel (1999). Innovation by lead users in a "second - tier" market: Patterns in product modification and information sharing. Working paper Sloan School of Management, MIT. Cambridge, MA.

Moscardo, G., P. Pearce, A. Morrison, D. Green and J. T. O'Leary (2000). "Developing a Typology for Understanding Visiting Friends and Relatives Markets." Journal of Travel Research **38**(3): 251 - 259.

Mosteller, F. and A. Parunak (1985). Identifying Extreme Cells in a Sizeable Contingency Table: Probabilistic and Exploratory Approaches. Exploring Data Tables, Trends and Shapes. D. C. Hoaglin, F. Mosteller and J. W. Tukey. New York, John Wiley & Sons Inc.

Muirhead, C. R. (1986). "Distinguishing Outlier Types in Time-Series." Journal of the Royal Statistical Society Series B-Methodological **48**(1): 39-47.

Mullen, M., G. R. Milne and P. M. Doney (1995). "An International Marketing Application of Outlier Analysis for Structural Equations: A Methodological Note." Journal of International Marketing **3**(1): 45 - 63.

Muncy, J. A. and R. P. Fisk (1987). "Cognitive Relativism and the Practice of Marketing Science." Journal of Marketing **51**(January): 20-23.

Mynatt, C., M. E. Doherty and R. D. Tweney (1978). "Consequences of confirmation and disconfirmation in a simulated research environment." Quarterly Journal of Experimental Psychology **30**: 395 - 406.

Naess, P. (2004). "Prediction, Regressions and Critical Realism." Journal of Critical Realism **3**(1): 133 - 164.

Nagel, E. (1961). The Structure of Science. New York, Hardcourt Brace Jovanovich.

Neal, W. D. (2000). Market Segmentation. Marketing Research: State-of-the-Art Perspectives. C. Chakrapani. Chicago, American Marketing Association**:** 375-399.

Newton, R. R. and K. E. Rudestam (1999). Your statistical consultant: Answers to your data analysis questions. Thousand Oaks, CA, Sage.

Neyman, J. and E. L. Scott (1971). Outlier Proneness of Phenomena and of Related Distributions. Optimizing Methods in Statistics. J. Rustagi. New York, Academia Press.

Nuts, R. and P. J. Rousseuw (1996). "Computing Depth Contours of Bivariate Point Clouds." Computational Statistics and Data Analysis: 153 - 168.

Oakes, M. (1986). Statistical Inference: A Commentary for the Social and Behavioural Sciences. New York, Wiley.

O'Donovan, D. (1965). "Rating Extremity: Pathology or Meaningfulness." Psychological Review **72**(5): 358 - 372.

Olsen, W. and J. A. Morgan (2005). "A Critical Epistemology of Analytical Statistics: Addressing the Sceptical Realist." Journal for the Theory of Social Behaviour 35(3): 255 - 284.

Olson, J. C. (1981). Towards a Science of Consumer Behavior. Advances in Consumer Research. A. A. Mitchell. Ann Arbor, Association for Consumer Research. **9**.

Onorato, M., A. R. Osborne and M. Serio (2002). "Extreme wave events in directional, random oceanic sea states." Physics of Fluids **14**(4): L25 - L28.

Oppenheim, A. N. (1966). Questionnaire design and attitude measurement. New York, Basic Books.

Oppermann, M. (1994). Comparative analysis of escorted tour packages in New Zealand and North America. Tourism Down Under: Tourism Research Conference, Department of Management Systems, Faculty of Business Studies, Massey University.

Orlov, A. I. (2005). "On Development of the Methodology of Statistical Methods." Journal of Mathematical Sciences **127**(4): 2119 - 2124.

Orr, J. M., P. R. Sackett and C. L. Z. Dubois (1991). "Outlier Detection and Treatment in I/O Psychology - a Survey of Researcher Beliefs and an Empirical Illustration." Personnel Psychology **44**(3): 473-486.

Osborne, J. W. (2002). "Notes on the use of data transformations." Practical Assessment Research and Evaluation **8**(6).

Osborne, J. W., W. R. I. Christiansen and J. S. Gunter (2001). Educational psychology from a statistician's perspective: A review of the quantitative quality of or field. Annual Meeting of the American Educational Research Association, Seattle, WA.

Osborne, J. W. and A. Overbay (2004). "The power of outliers (and why researchers should always check for them)." Practical Assessment Research and Evaluation **9**(6).

Outhwaite, W. (1987). New philosophies of social science: realism, hermeneutics and critical theory. New York.

Pan, G. W. and E. Laws (2002). "Tourism Marketing Opportunities for Australia in China." Journal of Vacation Marketing **8**(1): 39 - 48.

Papadimitriou, S. and C. Faloutsos (2003). "Cross-outlier detection." SSTD'03: 199 - 213.

Paulhus, D. L. (1984). "Two -Component Models of Socially Desirable Responding." Journal of Personality and Social Psychology **46**(3): 598-609.

Paulhus, D. L. (1991). Measurement and Control of Response Bias. Measures of Personality and Social Psychological Attitudes. J. P. Robinson, P. R. Shaver and L. S. Wright. San Diego, CA, Academic Press**:** 17 - 59.

Pearce, P. L. (2007). "A "University of Travel": Backpacker learning." Tourism Management **28**(5): 1285 - 1298.

Pena, D. and A. Maravall (1991). "Interpolation, Outliers and Inverse Autocorrelations." Communications in Statistics-Theory and Methods **20**(10): 3175-3186.

Penn, M. and E. K. Zalesne (2007). Microtrends: the small forces behind tomorrow's big changes. New York, Hachette Book Group.

Peter, J. P. and M. K. Olsen (1983). "Is Science Marketing." Journal of Marketing **47**: 111 - 125.

Peter, J. P., J. C. Olson and K. G. Grunert (1999). Consumer Behavior and Marketing Strategy. London, McGraw - Hill.

Petrick, J. F. (2004). "Are loyal visitors desired visitors?" Tourism Management **25**(4): 463-470.

Pettitt, L. I. (1988). "Bayes Methods for Outlier in Exponential Samples." Journal of the Royal Statistical Society Series B-Methodological **50**(3): 371 - 380.

Phillips, E. M. and D. S. Pugh (1994). How to Get a PhD: A Handbook for Students and their Supervisors. Buckingham, Open University Press.

Pimentel, R. W. and K. E. Reynolds (2004). "A Model for Consumer Devotion: Affective Commitment with Proactive Sustaining Behaviors." Academy of Marketing Science Review **2004**(5).

Platt, J. R. (1964). "Strong Inference." Science(146): 347 - 353.

Popper, K. (1959). The Logic of Scientific Discovery. New York, Harper.

Popper, K. (1962). Conjectures and Refutations. New York, Harper.

Popper, K. (1966). Logik der Forschung. Tübingen.

Popper, K. (1972). Objective Knowledge. Oxford, Clarendon.

Popper, K. (1989). Logik der Forschung. Tübingen, Mohr.

Porpora, D. (2001). Do Realists Run Regressions? After Postmodernism: An Introduction to Critical Realism. J. Lopez and G. Potter. New York, Athlone and Continuum Press.

Pratschke, J. (2003). "Realistic Models? Critical Realism and Statistical Models in the Social Sciences." Philosophica **71**: 13 - 38.

Prideaux, B. and H. McClymont (2006). "The changing profile of caravanners in Australia." International Journal of Tourism Research **8**(1): 45 - 58.

Proschan, F. (1953). "Rejection of Outlying Observations." American Journal of Physics **21**(7): 520-525.

Pruzek, R. M. (1997). An Introduction to Bayesian Inference and its Applications. What If There Were No Significance Tests? L. L. Harlow, S. A. Mulaik and J. H. Steiger. Mahwah, NJ, Lawrence Erlbaum Associates.

Psillos, S. (1999). Scientific Realism. London, Routledge.

Quine, W. V. (1953). From a Logical Point of View. Cambridge, MA, Harvard University Press.

Quinlan, J. R. (1993). C4.5 programs for machine learning. San Mateo, CA, Morgan Kaufmann Publishers.

Quinney, R. (1974). Critique of Legal Order: Crime, Control in Capitalistic Society. Boston, Little, Brown.

Quinney, R. (1975). Criminology. Boston, Little, Brown.

Quiroga, I. (1990). "Characteristics of Package Tours in Europe." Annals of Tourism Research **17**(2): 185-207.

Raghubir, P. and J. Srivastava (2002). "Effect of face value on product valuation in foreign currencies." Journal of Consumer Research **29**(3): 335-347.

Ramaswamy, S., R. Rastogi and S. Kyuseok (2000). "Efficient Algorithms for Mining Outliers from Large Data Sets." SIGMOD'00: 93 - 104.

Rasmussen, J. L. (1988). "Evaluating Outlier Identification Tests - Mahalandobis D Squared and Comrey Dk." Multivariate Behavioral Research **23**(2): 189 - 202.

Ray, J. J. (1983). "Reviving the Problem of Acquiescent Response Bias." Journal of Social Psychology **121**(1): 81-96.

Reichard, C. S. and S. F. Rallis (1994). The Qualitative-Quantitative Debate: New Perspectives. San Francicso, Jossey-Bass.

Reichenbach, H. (1938). Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge. Chicago, The University of Chicago Press.

Reiss, R. D. and M. Thomas (2001). Statistical Analysis of Extreme Values. Basel, Birkhäuser Verlag.

Richman, W. L., S. Kiesler, S. Weisband and F. Drasgow (1999). "A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews." Journal of Applied Psychology **84**(5): 754-775.

Riege, A. M. and C. Perry (2000). "National Marketing Strategies in International Travel and Tourism." European Journal of Marketing **34**(11/12): 1290 - 1305.

Ritchie, J. B. R., N. Carr and C. Cooper (2003). Managing Educational Tourism, Channel View Publications.

Robson, C. (2002). Real world research: a resource for social scientists and practitioner - researchers. Malden, MA, Blackwell Publishing.

Rosenthal, R. and R. L. Rosnow (1969). Artefact in behavioral research. New York, Academic Press.

Rosner, B. (1975). "Detection of Many Outliers." Technometrics **17**(2): 221-227.

Rosner, B. (1983). "Percentage Points for a Generalized ESD Many-Outlier Procedure." Technometrics **25**(2): 165-172.

Ross, S. M. (2003). "Peirce's criterion for the elimination of suspect experimental data." Journal of Engineering Technology(Fall).

Rossi, J. S. (1990). "Statistical Power of Psychological-Research - What Have We Gained in 20 Years." Journal of Consulting and Clinical Psychology **58**(5): 646-656.

Rothenberg, A. (1979). The Emerging Goddess: The Creative Process in Art, Science and Other Fields. New York, Norton.

Rothenberg, L. S. and M. S. Sanders (2000). "Lame duck politics: Impending departure and the votes on impeachment." Political Research Quarterly **53**: 523 - 536.

Roubach, M. (2004). "Meaning, phenomenology, and being." Inquiry(47): 189 - 199.

Rousseuw, P. J. and v. Zomeren (1987). Identification of Multivariate Outliers and Leverage Points by Means of Robust Covariance Matrices. Delft, Faculty of Mathematics and Informatics, Delft University of Technology.

Russel, B. (1961). A History of Western Philosophy. London, Allen & Unwil.

Savage, C. W. and P. Ehrlich (1991). Philosophical and Foundational Issues in Measurement Theory. New York, Lawrence Erlbaum Associates.

Sayer, A. (1992). Method in Social Science - A realist approach. London, Routledge.

Schmidt, F. L. and J. E. Hunter (1997). Eight Common but False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data. What If There Were No Significance Tests? L. L. Harlow, S. A. Mulaik and J. H. Steiger. Mahwah, NJ, Lawrence Erlbaum Associates.

Schmidt, W. (1986). "The Multiple Outlier Problem in Time Series Analysis." Australian Journal of Statistics **28**(3): 400 - 413.

Schofer, B. (2000). "Für einen moderaten Relativismus in der Wissenschaftssoziologie." Kölner Zeitschrift für Soziologie und Sozialpsychologie **52**(4): 696 - 719.

Scholkopf, B., J. R. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson (2001). "Estimating the support of a high dimensional distribution." Neural Computation **13**(7): 1443 - 1472.

Schuman, H. and S. Presser (1981). Questions and Answers in Attitude Surveys. New York, Academic Press.

Schwager, S. J. and B. H. Margolin (1982). "Detection of Multivariate Normal Outliers." Annals of Statistics **10**(3): 943-954.

Schweder, T. (1976). "Some Optimal Methods to Detect Structural Shift or Outliers in Regression." Journal of the American Statistical Association **71**(354): 491-501.

Schwertman, N. C., M. A. Owens and R. Adnan (2004). "A simple more general boxplot method for identifying outliers." Computational Statistics & Data Analysis **47**(1): 165-174.

Sharma, S., R. M. Durand and O. Gur-Arie (1981). "Identification and Analysis of Moderator Variables." Journal of Marketing Research **18**(August): 291 - 300.

Sharma, S. and V. Mahajan (1980). "Early Warning Indicators of Business Failure." Journal of Marketing **44**(4): 80 - 90.

Sherman, E. and E. S. Newman (1977). "The Meaning of Cherished Personal Possessions for the Elderly." Journal of Aging and Human Development **8**(2): 181 - 192.

Shugan, S. M. (2006). "Are Consumers Rational? Experimental Evidence?" Marketing Science **25**(1): 1-7.

hulman, A. (1973). "Comparison of 2 Scales on Extremity Response Bias." Public Opinion Quarterly **37**(3): 405-412.

Silverberg, K. E., S. J. Backman and K. F. Backman (1996). "A Preliminary Investigation into the Psychographics of Nature-Based Travelers to Southeastern United States." Journal of Travel Research **35**(2): 19 - 28.

Simon, D. R. (2002). Elite Deviance. Boston, Allyn and Bacon.

Simonoff, J. S. (1988). "Detecting Outlying Cells in 2-Way Contingency-Tables Via Backwards-Stepping." Technometrics **30**(3): 339-345.

Sinha, B. K. (1984). "Detection of Multivariate Outliers in Elliptically Symmetric Distributions." Annals of Statistics **12**(4): 1558-1565.

Sjö, E. (2001). "Simultaneous distributions of space-time wave characteristics in a Gaussian Sea." Extremes **4**: 263 - 288.

Slotegraaf, R. J., C. Moorman and J. Inman (2003). "The role of firm resources in returns to market deployment." Journal of Marketing Research **40**(3): 295-309.

Smart (1964). "The importance of negative results in psychological research." Canadian Psychologist **5**: 225 - 232.

Smith, M. C. and K. J. MacKay (2001). "The Organization of Information in Memory for Pictures of Tourist Destinations: Are there age related differences." Journal of Travel Research **39**(3): 261 - 266.

Smith, M. J. (1998). Social Research in Question. London / Milton Keynes, Open University Press.

Sneyers, R. (1984). Extremes in meteorology. Statistical Extremes and Applications. Dordrecht, The Netherlands, Reidel Publishing Company.

Solomon, M. R. and B. G. Englis (1994). "The Big Picture: Product Complementarity and Integrated Communications." Journal of Advertising Research **34**(6): 57 - 63.

Song, C., H. Shlomo and A. Makse (2005). "Self -Similarity of Complex Networks." Nature **433**: 392 - 395.

Stefansky, W. (1972). "Rejecting Outliers in Factorial Designs." Technometrics **14**(2): 469-&.

Steger, C. (2007). Segmentierung oder Individualisierung? Ein Vergleich von Produktstrategien zur Befriedigung heterogener Kundenbedürfnisse. Entrepreneurship and Innovation. Vienna, Vienna University of Economics and Business Administration.

Stening, B. W. and J. E. Everett (1984). "Response Styles in a Cross-Cultural Managerial Study." Journal of Social Psychology **122**(2): 151-156.

Stevens, J. P. (1992). Applied multivariate statistics for the social sciences. Hillsdale, NJ, Erlbaum.

Stolarski, r. S., A. J. Krueger, M. R. Schoeberl, R. D. McPeters, P. A. Newman and J. C. Alpert (1986). "Nimbus 7 satellite measurements of the springtime Antarctic ozone decrease." Nature **322**: 808 - 811.

Sutherland, E. H. (1939). Principles of Criminology. Philadelphia, Lippincott.

Tabachnick, B. G. and L. S. Findell (2000). Using multivariate statistics, Pearson Allyn & Bacon.

Tang, J., Z. Chen, A. W. Fu and D. W. Cheung (2002). "Enhancing Effectiveness of Outlier Detections for Low Density Patterns." PAKDD'02: 535 - 548.

Tax, D. M. J. and R. P. W. Duin (1999). "Support vector data description." Pattern Recognition Letters **20**(11-13): 1191 - 1199.

Tensaout, M. (2006). Généralisations empiriques en marketing: une revue critique. 12ième Congrès AFM, Nantes.

Thio, A. (2006). Deviant behavior. Boston, MA, Pearson.

Thio, A. (2009). Deviant beavhior, Boston, MA, Pearson

Thiruvengadam, A. (1972). Corrosion fatigue at high frequencies and high hydrostatic pressures. stress corrosion cracking of metal. A state of the art technical report. Philadelphia, ASTM.

Thomas, R. K. and R. Kirchner, J. (1991). Desktop Marketing. New York, American Demographic Books.

Thompson, B. (1996). "AERA editorial policies regarding statistical significance testing: Three suggested reforms." Educational Researcher **25**(26 - 30).

Thomson, C. and D. Pearce (1980). "Market Segmentation of New Zealand package tours." Journal of Travel Research **19**(2): 3 - 6.

Tietjen, G. and R. H. Moore (1972). "Some Grupps - Type Statistics for the Detection of Several Outliers." Technometrics **14**(3): 583 - 597.

Tsay, R. S. (1986). "Time Series Model Specification in the Presence of Outliers." Journal of the American Statistical Association **81**(393): 132 - 141.

Tsay, R. S. (1988). "Outliers, Level Shifts, and Variance Changes in Time-Series." Journal of Forecasting **7**(1): 1-20.

Tsay, R. S., D. Pena and A. E. Pankratz (2000). "Outliers in multivariate time series." Biometrika **87**(4): 789 - 804.

Tufte, E. R. (1983). The visual display of quantitative information. Cheshire, CT, Graphics Press.

Tufte, E. R. (1990). Envisioning Information. Cheshire, CT, Graphics Press.

Tukey, J. W. (1969). "Analyzing data: Sanctification or detective work?" American Psychologist **24**: 83 - 91.

Tukey, J. W. (1977). Exploratory Data Analysis. Princeton University, Addison - Wesely.

Twedt, W. (1964). "How Important to Marketing Strategy Is the ‘Heavy User’?" Journal of Marketing **28**(1): 71 - 73.

Uncles, M. and M. Wright (2004). "Empirical Generalizations in Marketing." Australasian Marketing Journal **12**(3): 5 - 29.

Underhill, P. (1999). Why we buy: The Science of Shopping. New York, Simon & Schuster.

Urban, G. L. and E. Von Hippel (1988). "Lead user analyses for the development of new industrial product." Management Science **34**(5): 569 - 582.

Van Middelkoop, M., A. Borgers and H. J. P. Timmermans (2003). "Inducing Heuristic Principles of Tourist Choice of Travel Mode: A Rule - Based Approach." Journal of Travel Research **42**(1): 75 - 83.

Van Selst, M. and P. Jolicoeur (1994). "A solution to the effect of sample size on outlier elimination." The Quarterly Journal of Experimental psychology **47**(3): 631 - 650.

Von Hippel, E. (1988). The sources of innovation. New York, Oxford University Press.

Wainer, H. (1976). "Robust Statistics: A survey and some prescriptions." Journal of Educational Statistics **1**(4): 285 - 312.

Wallendorf, M. and E. J. Arnould (1988). "My Favorite Things - a Cross-Cultural Inquiry into Object Attachment, Possessiveness, and Social Linkage." Journal of Consumer Research **14**(4): 531-547.

Watkins, D. and S. Cheung (1995). "Culture, Gender, and Response Bias - an Analysis of Responses to the Self-Description Questionnaire." Journal of Cross-Cultural Psychology **26**(5): 490-504.

Watson, D. (1992). "Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale - an Application to Class-Consciousness." Sociological Methods & Research **21**(1): 52-88.

Watts, D. (2003). Six Degrees: The Science of a Connected Age. New York, Norton.

Webster, C. (1996). "Hispanic and Anglo interviewer and Respondent Ethnicity and Gender: The impact on survey response quality." Journal of Marketing Research **33**(1): 62-72.

Wedel, M. and W. A. Kamakura (2000). Market Segmentation: Conceptual and Methodological Foundations. Boston, Kluwer Academic Publishers.

Wedel, M., W. A. Kamakura and U. Böckenholt (2000). "Marketing Data, Models and Decisions." International Journal of Research in Marketing **17**(2): 203 - 208.

Wei, L., W. Qian, A. Zhou, W. Jin and J. X. Yu (2003). "HOT: Hypergraph-Based Outlier Test for Categorical Data." PAKDD'03: 399 - 410.

Weick, K. E. and K. M. Sutcliffe (2001). Managing the Unexpected: Assuring High Performance in an Age of Complexity. San Francisco, Jossey-Bass.

Wells, W. D. (1963). "How Chronic Overclaimers Distort Survey Findings." Journal of Advertising Research **3**(2): 8-18.

West, M. (1984). "Outlier Models and Prior Distributions in Bayesian Linear-Regression." Journal of the Royal Statistical Society Series B-Methodological **46**(3): 431-439.

Wijnen, K., I. Vermeir and P. Van Kenhove (2007). "The relationships between traits, personal values, topic involvement, and topic sensitivity in a mail survey context." Personality and Individual Differences **42**(1): 61-73.

Wilcox, R. R. (1998). "How many discoveries have been lost by ignoring modern statistical methods?" American Psychologist **53**(3): 300-314.

Wilkinson, L. (1999). "Statistical Methods in Psychology Journals - Guidelines and Explanations." American Psychologist **54**(8): 594 - 604.

Williams, G. J., R. A. Baster, H. He, S. Harkins and L. Gu (2002). "A Comparative Study of RNN for Outlier Detection in Data Mining." ICDM'02: 709 - 712.

Williams, M. and W. Dyer (2004). Realism and Probability. Making Realism Work. B. Carter and C. New. Oxfordshire, Routledge.

Willis, G. B. (2005). Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA, Sage.

Wills, J., A. C. Samli and L. W. Jacobs (1991). "Developing Global Products and Marketing Strategies: A Construct and a research Agenda." Journal of the Academy of Marketing Science **19**(1): 1 - 10.

Wilson, F. D., G. L. Smoke and J. D. Martin (1973). "The replication problem in sociology: A Report and a Suggestion." Sociological Inquiry **43**: 141 - 149.

Wind, Y. (1977). "Toward a Change in the Focus of Marketing Analysis: From a Single Brand to an Assortment." Journal of Marketing **12**(4): 143.

Winkler, J. D., D. E. Kanouse and J. E. Ware (1982). "Controlling for Acquiescence Response Set in Scale Development." Journal of Applied Psychology **67**(5): 555-561.

Woodside, A. G. and C. Dubelaar (2002). "A General Theory of Tourism Consumption Systems: A Conceptual Framework and an Empirical Exploration." Journal of Travel Research **41**(November): 120 - 132.

Wright, B. R. E., A. Caspi, T. E. Moffitt and R. Paternoster (2004). "Does the perceived risk of punishment deter criminally prone individuals? Rational choice, self - control, and crime." Journal of Research in Crime and Delinquency **41**: 180 - 213.

Wyatt, D. F. and D. T. Campbell (1951). "On the liability of stereotype or hypothesis." Journal of Abnormal and Social Psychology **46**: 496 - 500.

Wyer, R. S. (1969). "Effects of General Response Style on Measurement of Own Attitude and Interpretation of Attitude-Relevant Messages." British Journal of Social and Clinical Psychology **8**(JUN): 104-115.

Yamanishi, K. and J. Takeuchi (2001). "Discovering Outlier filtering Rules from Unlabeled Data - Combining a Supervised Learner with an Unsupervised Learner." KDD'01: 389 - 394.

Yamanishi, K., J. Takeuchi and G. Williams (2000). "On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms." KDD'00: 320 - 325.

Yu, D., G. Sheikholeslami and A. Zhang (2002). "FindOut: Finding Out Outliers in Large Datasets." Knowledge and Information Systems **4**(4): 387 - 412.

Zimmerman, D. W. (1994). "A Note on the Influence of Outliers on Parametric and Nonparametric - Tests." Journal of General Psychology **121**(4): 391 - 401.

Zimmerman, D. W. (1995a). "Increasing the power of nonparametric tests by detecting and downweighting outliers." Journal of Experimental Education **64**(1): 71-78.

Zimmerman, D. W. (1995b). "Increasing the Power of the Anova-F Test for Outlier-Prone Distributions by Modified Ranking Methods." Journal of General Psychology **122**(1): 83-8