TECHNIQUES IN THE BEHAVIORAL AND NEURAL SCIENCES
VOLUME 14

Series editor: J.P. Huston

EXPERIMENTAL DESIGN
A HANDBOOK AND DICTIONARY FOR MEDICAL AND BEHAVIORAL RESEARCH

J. Krauth

		Phase 2	
		Treatment A	Treatment B
Phase 1	Treatment A	A_1A_2	A_1B_2
	Treatment B	B_1A_2	B_1B_2

ELSEVIER

TECHNIQUES IN THE BEHAVIORAL AND NEURAL SCIENCES

VOLUME 14

EXPERIMENTAL DESIGN. A HANDBOOK AND DICTIONARY FOR MEDICAL AND BEHAVIORAL RESEARCH

Previously published in
TECHNIQUES IN THE BEHAVIORAL AND NEURAL SCIENCES

Volume 1: Feeding and Drinking, by F. Toates and N.E. Rowland (Eds), 1987, ISBN 0-444-80895-7

Volume 2: Distribution-Free Statistics: An Application-Oriented Approach, by J. Krauth, 1988, ISBN 0-444-80934-1, Paperback ISBN 0-444-80988-0

Volume 3: Molecular Neuroanatomy, by F.W. Van Leeuwen, R.M. Buijs, C.W. Pool and O. Pach (Eds), 1989, ISBN 0-444-81014-5, Paperback ISBN 0-444-81016-1

Volume 4: Manual of Microsurgery on the Laboratory Rat, Part 1, by J.J. van Dongen, R. Remie, J.W. Rensema and G.H.J. van Wunnik (Eds), 1990, ISBN 0-444-81138-9, Paperback ISBN 0-444-81139-7

Volume 5: Digital Biosignal Processing, by R. Weitkunat (Ed), 1991, ISBN 0-444-81140-0, Paperback ISBN 0-444-98144-7

Volume 6: Experimental Analysis of Behavior, by I.H. Iversen and K.A. Lattal (Eds), 1991, Part 1, ISBN 0-444-81251-2, Paperback ISBN 0-444-89160-9, Part 2, ISBN 0-444-89194-3, Paperback ISBN 0-444-89195-1

Volume 7: Microdialysis in the Neurosciences, by T.E. Robinson and J.B. Justice Jr. (Eds), 1991, ISBN 0-444-81194-X, Paperback ISBN 0-444-89375-X

Volume 8: Techniques for the Genetic Analysis of Brain and Behavior, by D. Goldowitz, D. Wahlsten and R.E. Wimer (Eds), 1992, ISBN 0-444-81249-0, Paperback ISBN 0-444-89682- 1

Volume 9: Research Designs and Methods in Psychiatry, by M. Fava and J.F. Rosenbaum (Eds), 1992, ISBN 0-444-89595-7, Paperback ISBN 0-444-89594-9

Volume 10: Methods in Behavioral Pharmacology, by F. van Haaren (Ed), 1993, ISBN 0-444-81444-2, Paperback ISBN 0-444-81445-0

Volume 11: Methods in Neurotransmitter and Neuropeptide Research, by S.H. Parvez, M. Naoi, T. Nagatsu and S. Parvez (Eds), 1993, Part 1, ISBN 0-444-81369-1, Paperback ISBN 0-444-81674-7, Part 2, ISBN 0-444-81368-3, Paperback ISBN 0-444-81675-5

Volume 12: Neglected Factors in Pharmacology and Neuroscience Research, by V. Claassen, 1994, ISBN 0-444-81871-5, Paperback ISBN 0-444-81907-X

Volume 13: Handbook of Molecular-Genetic Techniques for Brain and Behavior Research, by W.E. Crusio and R.T. Gerlai (Eds), 1999, ISBN 0-444-50239-4

Techniques in the Behavioral and Neural
Sciences

Series Editor

J.P. HUSTON
Düsseldorf

Volume 14

EXPERIMENTAL DESIGN
A HANDBOOK AND DICTIONARY FOR MEDICAL AND BEHAVIORAL RESEARCH

by

J. KRAUTH
*Heinrich-Heine-Universität, Psychologisches Institut, Lehrstuhl IV,
Universitätsstrasse 1, D-40225 Düsseldorf, Germany*

2000

ELSEVIER
AMSTERDAM - LONDON - NEW YORK - OXFORD - PARIS - SHANNON - TOKYO

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

Notice
No responsibility is assumed by the Publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

First edition 2000

Library of Congress Cataloging in Publication Data
A catalog record from the Library of Congress has been applied for.

ISBN: 0-444-50637-3 (Hardbound)
ISBN: 0-444-50638-1 (Paperback)
ISSN Series: 0921-0709

♾ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).
Printed in The Netherlands.

Preface

Numerous books treating the planning and conduction of experiments and studies have been published in the past. In fact every empirical researcher needs sufficient knowledge in this field, as there is no other way to get the data he or she needs to answer his or her questions. Everyone who is at least to some degree experienced in experimental design is aware of the errors he has to avoid when planning his own research and will also easily detect the errors in the studies done by others. For such a researcher experimental design is nothing but the application of some simple basic rules suggested by common sense. However, one ought to keep in mind that this profound knowledge has been derived from a long grown experience with the interpretation of empirical data. Of course, the mere study of a book about experimental design cannot make up for the fundamental knowledge gained over many years by trial and error. The present handbook tries to display at least the basic rules of a rational experimental design in medical and behavioral research. A summary and questions are given at the end of each chapter; the respective answers can be found after Chapter 9.

In what regard does the present handbook differ from other publications on the same subject? First, it is restricted to medical and behavioral research. This implies that, unlike in e.g. physics, chemistry, or agriculture, the planning of studies dealing with the behavior of organisms, i.e., human beings or animals, is considered. Whether this is de facto a restriction might be disputed, as the basic principles of such a planning are the same. Furthermore, many more aspects must be paid attention to, if one investigates behavior as opposed to, e.g., the effectiveness of fertilizers.

A second difference of the present handbook if compared to other books is its restriction to the principles of the planning of studies, methods of data analysis in general and of statistics in particular are not discussed. This restriction is made on the one hand, because it seemed to be justified to dispense with the description of statistical methods in a book on experimental design. On the other hand, an appropriate presentation of methods of data analysis would have considerably inflated the size of this book. Finally, many books on data analysis and statistics exist and may be used in addition to this handbook. However, there is no denying the fact that a separate discussion of the design and analysis of experiments appears to be rather artificial. Many authors thus try a synthesis of both aspects. A rather weak argument in favor of our approach might be that it makes, at least, more sense to describe the principles of experimental design without the corresponding data analysis than doing it the other way round. However, we have to admit that a reader who is not familiar at all with statistical reasoning may experience some difficulties in really understanding certain sections of this book (e.g. Sections 3.1.1, 3.1.2, 3.1.3, 7.3, 9.2.1, 9.2.2, and 9.4). These sections are of no importance with respect to the comprehension of the principles of experimental research and may be skipped by readers who have not had an introduction into the theory of statistical tests or, with respect to Section 7.3, into the analysis of variance. We included this material for those readers who have at least had an introductory course in statistics, and may profit from this additional information.

A third difference refers to the omission of the philosophical foundations of experimental research in general and of experimental design in particular. Many authors regard the description of the underlying philosophy as the appropriate way to introduce the principles of experimental design. Indeed, leaving out these foundations might seem daring. The present book, however, is intended for readers who are familiar with these foundations, who are not interested in the philosophical considerations, or who are simply convinced that the experimental approach is the only valid scientific strategy.

A fourth difference in comparison to other books is the emphasis we lay on the importance of the principle of randomization and our disapproval of repeated-measures designs. Of course, both topics are also being discussed in nearly any other book on experimental design. Nevertheless, readers of these books often seem to get the impression that it is possible to draw causal conclusions from studies without an appropriate randomization or that the use of sophisticated statistical procedures guarantees that a conclusive interpretation of repeated-measures designs is possible. It is a primary concern of this book to make the reader aware of this misconception. However, this does not mean that repeated-measures designs are not thoroughly discussed. On the contrary, a considerable part of the book is devoted to the discussion of these designs (cf., e.g., Sections 4.8, 4.12, 6.3, Chapters 8 and 9). In particular, we explain, in which way designs with repeated measures are to be constructed to allow for causal conclusions. This point may surprise readers with a certain background in statistics because statistical procedures for evaluating classical repeated-measures designs have been in use for many years in many fields of science. Here, in most cases the problem is not in the first place the validity of the statistical procedure which is used, but the substantial interpretation of the results gained by this procedure. If, e.g., a researcher performs a pretest in a group of subjects, then a treatment, and then a posttest, the pretest and posttest values may be compared by means of a paired t-test. One might conclude from a significant result if the assumptions for the paired t-test are met that the pretest and posttest values have a different distribution. However, it is not allowed to conclude that this is an effect of the treatment because without a control group we cannot rule out that we would have observed the same effect if no treatment had been applied.

The importance which we assign to the principle of randomization may seem exaggerated if we take the many epidemiological and other field studies, where randomization is not feasible, into account. However, as we shall see, the outcomes of such non-randomized studies never permit an unequivocal interpretation, even if other experimental principles (control groups, matching etc.) have been used.

Differing even further from other books, the design of experiments in medical as well as in behavioral research will be treated simultaneously. Though in many cases the designs are similar or even identical for the two fields, various characteristic features exist, which are typical of the respective field. Thus, here, both medical and behavioral researchers will find answers to their questions.

Finally, a last point deals with the general structure of this book. Some authors have tried to provide a systematic presentation of experimental design. Certainly, this approach does have its advantages: there are no gaps and, having understood the system, the reader will easily find or even construct the designs he or she needs. Since there are infinitely many possible designs which can be constructed on the basis of only a handful of basic principles, this kind of systematic presentation suggests itself and will appeal to those readers who like a systematic approach. Other readers, however, might refuse the idea of having to understand a specific systematic ordering of the designs first, which has been invented by an author, before being able to identify those designs which they are interested in. It might suit these readers much more, if particular subclasses of designs are combined following a certain principle. Here, overlappings and omissions cannot be avoided, though this is also true, to a certain degree, for the systematic approach. Though we do not give a systematic listing of designs, we have tried to consider as many aspects of experimental design as possible in this handbook. For those readers who do not want to use this book as a systematic introduction, but rather as a reference book, the subject index and, in particular, the "Dictionary of Experimental Design" will be of use.

Of the many books on experimental design which have inspired me there are three books which impressed me in particular by their originality and their organization. These books are Frank J. McGuigan's "Experimental Psychology", "Introduction to Experimental Psychology" by Douglas W. Matheson, Richard L. Bruce, and Kenneth L. Beauchamp, and "Quasi-Experimentation" by Thomas D. Cook and Donald T. Campbell. Though I may differ from those authors in more than one respect, their reflections were of great value to me when composing this handbook. As far as the "Dictionary of Experimental Design" is concerned, "Elsevier's Dictionary of Biometry" by Dieter Rasch, Moti Lal Tiku and Dieter Sumpf, "The Cambridge Dictionary of Statistics in the Medical Sciences" by Brian S. Everitt, and the "Dictionary of Statistics and Methodology" by W. Paul Vogt provided many interesting ideas.

I have to thank Frank Wesselmann for transforming my manuscript into a legible form and Janine Illian for smoothing my English.

Joachim Krauth

Contents

Preface v

A. Handbook of Experimental Design 1

1. Historical Remarks 2

1.1 The Diet Experiment of the Prophet Daniel 2
1.2 The Lemon Experiment of an Egyptian Judge 3
1.3 Drug Research in the 11th Century 3
1.4 John Stuart Mill and the Foundations of Experimental Research 4
1.5 Wilhelm Wundt and the Experiment in Psychology 6
1.6 The Invention of Randomization 7
1.7 Sir Ronald Fisher and Randomization 10

Summary 13

Questions 14

2. The Object of Experimental Design 15

2.1 Dependent and Independent Variables 15
2.2 Selection of Factor Levels 15
2.3 Causal Relations and Intervening Variables 16
2.4 Ockham's Razor 17
2.5 Constructs 18
2.6 Causal and Correlative Relations 19

Summary 20

Questions 20

3. A Case for Experimental Design 21

3.1 Threats to Statistical Conclusion Validity 22

3.1.1 Low Statistical Power 22
3.1.2 Violated Assumptions of Statistical Tests 23
3.1.3 Multiple Tests 24
3.1.4 Reliability of the Dependent Variables 24
3.1.5 Reliability of the Independent Variables 25
3.1.6 Random Disturbance of the Experimental Situation 25
3.1.7 Random Differences between Subjects 25

3.2 Threats to Internal Validity 25

3.2.1 History 25
3.2.2 Maturation 26

3.2.3 Testing 26
3.2.4 Instrumentation 26
3.2.5 Statistical Regression 26
3.2.6 Selection 28
3.2.7 Experimental Mortality 28
3.2.8 Direction of the Causal Conclusion 28
3.2.9 Exchange of Information 29

3.3 Threats to Construct Validity 29

3.3.1 Inexact Definitions of Constructs 29
3.3.2 Mono-Operation Bias 29
3.3.3 Mono-Method Bias 30
3.3.4 Hypothesis Guessing 30
3.3.5 Social Desirability Responding 30
3.3.6 Experimenter Expectancies 30
3.3.7 Clever Hans and his Friends 31
3.3.8 Omitting Relevant Levels of Constructs 32
3.3.9 Effects of more than One Independent Variable 32
3.3.10 Interaction of Testing and Treatment 32
3.3.11 Restricted Generalizability over the Constructs 33

3.4 Threats to External Validity 33

3.4.1 Interaction between Selection and Treatment 33
3.4.2 Interaction between Setting and Treatment 34
3.4.3 Interaction between History and Treatment 34

Summary 35

Questions 35

4. Control of Extraneous Variables 36

4.1 Randomization 36

4.1.1 Randomization and External Validity 36
4.1.2 Randomization and Internal Validity 39
4.1.3 Randomization in Factorial Designs 40
4.1.4 Multiple Treatments 40
4.1.5 Randomization and Ethics 45

4.2 Elimination and Blocking Off 48
4.3 Constancy and Covering 49
4.4 Matching and Blocking 50
4.5 Extraneous Variables as Independent Variables 55
4.6 Replication 56
4.7 Balancing 58
4.8 Counterbalancing 60
4.9 Blinding 65
4.10 Control Groups and Control Conditions 67

4.10.1 Control Groups for Detecting Effects 67
4.10.2 Placebo Control Groups 68
4.10.3 Control Groups for the Case that no Effects can be Detected 70
4.10.4 Yoked Control Groups 71
4.10.5 Expectancy Control Groups 73
4.10.6 Solomon Design 76
4.10.7 Comparison Groups 78
4.10.8 Historical Controls 79

4.11 Conservative Arrangement of the Levels of Extraneous Variables 79
4.12 Repeated Measures 82
4.13 Statistical Adjustment 85

Summary 87

Questions 87

5. Preliminary Experiments and Pilot Studies 89

Summary 91

Questions 91

6. Designs which had Better be Avoided 92

6.1 Designs without Randomization 92
6.2 Designs without a Control Group 93
6.3 Designs with Repeated Measures 94
6.4 Crossover Designs 95
6.5 Designs with more than Two Factors 99

Summary 106

Questions 106

7. Designs without Repeated Measures 109

7.1 Designs with One Independent Variable 109
7.2 Designs with Two Independent Variables 112
7.3 Designs with more than Two Independent Variables 115

Summary 122

Questions 122

8. Designs with Repeated Measures 124

8.1 Designs with One Independent Variable 124
8.2 Designs with more than One Independent Variable 135

Summary 140

Questions 140

9. Single-Case Experimental Designs 141

9.1 Basic Principles of Single-Case Experimental Designs 141
9.2 Selected Single-Case Experimental Designs 144

9.2.1 Design with only One Factor with only Two Levels 144
9.2.2 Interrupted Time-Series Design 145

9.3 An Alternative Principle of Single-Case Experimental Designs 145
9.4 Combination of the Results of Several Independent Single-Case Experimental Designs 146

Summary 149

Questions 149

Answers to the Questions (with References) 151

B. Dictionary of Experimental Design 209

References 275

Author Index 281

Subject Index 283

This Page Intentionally Left Blank

PART A

Handbook of Experimental Design

1 Historical Remarks

The two most vital principles needed to appropriately plan experiments are the use of **control groups** or **control conditions** on the one hand, and the use of **randomization**, i.e. a random selection or a random assignment of individuals, on the other hand. The first principle has been known for at least two thousand years. This illustrates that at the same time when Aristotle and other Greek philosophers used a non-empirical approach an experimental approach might already have existed, accepting only statements based on empirical evidence gained in a systematic way. It is remarkable that the second principle of experimental design, i.e. randomization, which nowadays is assumed to be as important as the first principle, was invented at the end of the nineteenth century only and seems to have not been known in former times.

In the following short presentation only some sources will be presented in a rather anecdotal manner, whereas other sources will not even be mentioned, like e.g. certain statements of the Greek physician Galen or Galenius (129-199) or of the English politician and philosopher Francis Bacon (1561-1626).

1.1 The Diet Experiment of the Prophet Daniel

To begin with a quote from the book Daniel (Daniel 1, 10-13) of the Old Testament, which is said to date back to the time of the Maccabeens (167-164 B. C.), is given. There we find (Luther, 1862, p. 750):

> "(10) Derselbe sprach zu ihm: Ich fürchte mich vor meinem Herrn, dem Könige, der euch eure Speise und Trank verschafft hat, wo er würde sehen, daß eure Angesichter jämmerlicher wären, denn der andern Knaben eures Alters, so brächtet ihr mich bei dem Könige um mein Leben.
>
> (11) Da sprach Daniel zu Melzar, welchem der oberste Kämmerer Daniel, Hananja, Misael und Asarja befohlen hatte:
>
> (12) Versuche es doch mit deinen Knechten zehn Tage, und laß uns geben Zugemüse zu essen und Wasser zu trinken.
>
> (13) Und laß dann vor dir unsere Gestalt und der Knaben, so von des Königs Speise essen, besehen; und darnach du sehen wirst, darnach schaffe mit deinen Knechten."

In our translation this reads:

> "(10) The same said to him: I am afraid of my lord, the king, who has given you to eat and to drink. If he saw that your faces are more pitiable than those of other boys of your age, it would cost my life.
>
> (11) Daniel said to Melzar, to whom the first chamberlain had entrusted Daniel, Hananiah, Mishael and Azariah:
>
> (12) Just try it with your subjects ten days and let us eat vegetables and drink water.
>
> (13) After this time look at our figures and at those of the boys who eat the king's food. According to what you see you may decide furtheron about your subjects."

Thus, Daniel proposes that the four young Israelite hostages at the court of the Babylonian king Nebuchadnezar form an **experimental group** and a group of boys of the same age from the king's court a **control group**. The comparison of the two groups was supposed to allow a conclusion about whether a vegetarian diet without alcohol lets the young men appear more healthy and well-fed than does the usual food at the court.

1.2 The Lemon Experiment of an Egyptian Judge

The second example is taken from the book "Deipnosophistae" (The Banquet of the Sophists) by the Greek author Athenaeus or Athenaios from Naukratis in Egypt. This book is probably the oldest cookery book known today. Though the exact dates of birth and death of Athenaeus are unknown, one might conclude from certain statements in the opus that the author lived at the end of the second and at the beginning of the third century in Rome and that the book was written only some years after the year 228. In Volume III of the book (part 84-85) we find (Athenaeus, 1971, p. 365) that an Egyptian judge sentenced several criminals to death. The condemned persons were to be put to death by the bites of venomous asps in the theatre. On their way to the execution a peddler had pity on the convicts and offered them some pieces of lemon to eat. To the surprise of the judge they survived the bites of the asps. When the judge learned about the lemons, he again had two convicts, of whom one had gotten a piece of lemon while the other had not, bitten by the asps the next day. The convict who had not eaten a lemon died at once, the other one survived.

Here, again, the principle of using a **control group** and, in addition, a further principle, the principle of **replication**, as the experiment is said to have been repeated several times, is being depicted. Note by the way, that Athenaeus concluded from this report that lemons are an antidote to all kinds of poisons.

1.3 Drug Research in the 11th Century

Avicenna or in Arabian Ibn Sina (980-1037) was a Persian physician and philosopher who wrote numerous books on medicine and philosophy. Here, we are particularly interested in his opus Canon of Medicine (al-Quanum fi at-tibb) which according to Shah (1966) was used as a medical textbook by the universities Saint Louis and Montpellier until 1657. This means that this opus served as a textbook for more than 600 years! The second volume of a whole of five volumes treats simple drugs. While the second part of this volume lists the properties and applications of 760 different drugs, the first part describes the basic principles of experimental drug research. For that purpose, Avicenna formulated the following seven rules (cf. the Latin text in Crombie, 1952, pp. 103-104):

> **Rule 1** states that a remedy ought to be free of irrelevant characteristics. If we consider, e.g., heated water, it is not possible to distinguish whether the water or the heat has had an effect. Nowadays we would argue that one should avoid situations where several possible causal variables are effective at the same time, because then, one cannot conclude which of the causal variables has caused the observed effect.
>
> **Rule 2** states that the effect of a remedy should be tested on patients who suffer from a simple and not from a composite malady. Otherwise it is not possible to identify the real

cause of a recovery. Therefore, effect variables should be selected in such a way that the site of the effect can be identified unequivocally.

Rule 3 states that it is not sufficient to test a drug with only one type of a disease, because the drug may show an effect due to a particular situation. Nowadays we would argue that the effect of a drug can only be established if control groups are used.

Rule 4 states that the effect of a drug should be tested for different degrees of illness. Nowadays we would demand to establish a dose-response curve. Here, in contrast to Avicenna, we would test different doses of the drug for the same degree of illness.

Rule 5 requires an exact protocol of the experiment, since otherwise the real cause for a recovery cannot be separated from other possible causes. Another possible interpretation of this rule could be that causal variables, e.g. drugs, should not be altered during the experiment.

Rule 6 requires that the positive effect of a drug should always or at least in many cases be observed, because otherwise it may be a random effect. Here, Avicenna addresses the principle of replication of experiments. This principle says that the existence of an effect can only be maintained if this effect can be replicated with a high reliability.

Finally, **rule 7** requires that a drug should be tested with human beings and not, e.g., with lions or horses. Otherwise a positive effect cannot be taken for granted. This is in accord with the modern view that effects found for one species should not be generalized without more ado to another species.

1.4 John Stuart Mill and the Foundations of Experimental Research

The English philosopher, economist, historian and politician John Stuart Mill (1806-1873) published his opus "System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, And the Methods of Scientific Investigation" in 1843. It consists of two volumes; we cite here from the second edition which appeared in 1846. Book III, which is entitled "Of Induction", is of particular interest to our topic. Though the author discusses the possibilities of interpreting the results of experiments throughout the entire Book III, we will only consider Chapter VII "Of Observation and Experiment" and Chapter VIII "Of the Four Methods of Experimental Inquiry" in this context.

In Chapter VII experiment and observational study are compared as well in §3 "Advantages of experiment over observation" as in §4 "Advantages of observation over experiment". In the latter Mill also emphasizes, in spite of the title, that causal conclusions are only possible on the basis of experiments (Mill, 1846, pp. 447-448, italics by Mill):

> "If we can produce the antecedent artificially, and if, when we do so, the effect follows, the induction is complete; that antecedent is the cause of that consequent*. But we then have added the evidence of experiment to that of simple observation. Until we had done so, we had only proved *invariable* antecedence, but not *unconditional* antecedence, or causation. Until it had been shown by the actual production of the antecedent under

> known circumstances, and the occurrence thereupon of the consequent, that the antecedent was really the condition on which it depended; the uniformity of succession which was proved to exist between them might, for aught we knew, be (like the succession of day and night) no case of causation at all; both antecedent and consequent might be successive stages of the effect of an ulterior cause. Observation, in short, without experiment (and without any aid from deduction) can ascertain uniformities, but cannot prove causation."

In a footnote Mill (1846, p. 448) points out that the above inference is not really compelling:

> "*Unless, indeed, the consequent was generated not by the antecedent, but by the means we employed to produce the antecedent. As, however, these means are under our power, there is so far a probability that they are also sufficiently within our knowledge, to enable us to judge whether that could be the case or not."

In summary, Mill argues, that experiments, as opposed to observational studies, are planned in such a way that, at least in principle, causal conclusions are possible.

In Chapter VIII ("Of the Four Methods of Experimental Inquiry"), Mill describes, slightly contradicting the title, five methods of experimental research. The basic principle of each method is summarized in a canon, which we will quote in the sequel.

For the first method (**Method of Agreement**) the following canon is formulated (Mill, 1864, p. 454):

> *"If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree, is the cause (or effect) of the given phenomenon."*

For the second method (**Method of Difference**) we find (Mill, 1864, p. 455):

> *"If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon."*

The third method is called **Indirect Method of Difference** or **Joint Method of Agreement and Difference** (Mill, 1864, p. 463):

> *"If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance; the circumstance in which alone the two sets of instances differ, is the effect, or cause, or a necessary part of the cause, of the phenomenon."*

The canon of the fourth method (**Method of Residues**) is given by (Mill, 1864, p. 465):

> *"Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents."*

The fifth and final method (**Method of Concomitant Variations**) is characterized by the canon (Mill, 1864, p. 470):

> *"Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation."*

In the following Chapter IX ("Miscellaneous Examples of the Four Methods"), Mill uses a number of examples from the natural sciences to illustrate how the five methods may be used for drawing causal conclusions from experimental results.

Nevertheless, Mill never describes explicitly, how experiments are to be conducted to allow causal conclusions. He only discusses the methods by which, in certain situations, conclusions might be drawn from the given information. This kind of reasoning may be of use in testing whether given experimental designs are suited at all for drawing causal conclusions.

1.5 Wilhelm Wundt and the Experiment in Psychology

Modern experimental Psychology began in 1879 when the physiologist Wilhelm Max Wundt (1832-1920) founded the first psychological laboratory in Leipzig, who therefore can be considered as the first experimental psychologist. Wundt defined an experiment (Wundt, 1911, p. 25, spacing by Wundt) as follows:

> "Das Experiment besteht in einer Beobachtung, die sich mit der willkürlichen Einwirkung des Beobachters auf die Entstehung und den Verlauf der zu beobachtenden Erscheinungen verbindet. Die Beobachtung im engeren Sinn untersucht die Erscheinungen ohne derartige Einwirkungen, so wie sie sich in dem Zusammenhang der Erfahrung von selbst dem Beobachter darbieten. Wo überhaupt eine experimentelle Einwirkung möglich ist, da pflegt man diese in der Naturwissenschaft stets anzuwenden, weil es unter allen Umständen, auch wenn die Erscheinungen an und für sich schon einer zureichend exakten Beobachtung zugänglich sind, von Vorteil ist, Eintritt und Verlauf derselben willkürlich bestimmen oder auch einzelne Teile einer zusammengesetzten Erscheinung willkürlich isolieren zu können."

In our translation this reads:

> "An experiment consists of an observation which is connected with the observer's arbitrary manipulation of the formation and course of the phenomena which are to be observed. The observation in a restricted sense studies the phenomena without such manipulations, just as they are met by the observer. The experimental manipulation is used in the natural sciences wherever it is possible, as it is always advantageous to arbitrarily determine the beginning and the course of the phenomena or to be able to isolate parts of a more complex phenomenon in an arbitrary way. This holds even if the phenomena can be observed in an exact way without any manipulations."

From this citation the following principles or requirements or advantages of experiments when compared to observational studies can be extracted:

1. Manipulation

The experimenter causes the formation of the phenomena to be studied and determines their course.

Manipulation does not imply that the experimenter determines the outcome of the experiment in advance; he or she only fixes the frame or setting of the experiment.

2. Arbitrariness

The experimenter arbitrarily fixes the timing of the phenomenon of interest as well as the times of interventions, which may influence its course. The type of intervention itself is also determined arbitrarily by the experimenter.

Note, that in this context arbitrariness does not mean that any unreasonable intervention schemes or intervention methods are used. Rather, external conditions should not have any systematic influence on the outcome of experiments.

3. Reproducibility

An experiment may be repeated in the same way at any time. Reproducibility is a consequence of manipulation and arbitrariness.

4. Isolation

The experimenter may isolate a single causal variable from a complex of such variables and study it separately. Thus an observed effect can unequivocally be ascribed to a particular cause.

1.6 The Invention of Randomization

As already mentioned at the beginning of Chapter 1, a fundamental and indispensable technique for experimental design, called **randomization**, was first proposed at the end of the nineteenth century. The invention of this technique is usually ascribed to R. A. Fisher (1925), as discussed in the next section. This is, at least partly, due to a rather controversial discussion initiated by Fisher's suggestion which eventually led to the general acceptance of the technique in experimental design. Fisher (1925) does not give any reference to former authors who had proposed or used this technique before. While the term "randomization" itself seems to have been introduced by Fisher, we found two references from before 1925, where a randomization procedure was proposed or applied though the exact term was not mentioned. However, we cannot rule out the possibility that even earlier references exist.

The first reference known to us is an article by Charles S. Peirce and Joseph Jastrow (1885) in the tradition of Fechner (1860), dealing with small differences of sensation. The article is based on the well-known opus "Elemente der Psychophysik (Elements of Psychophysics)" (Fechner, 1860) by the physicist and philosopher Gustav Theodor Fechner (1801-1887). Fechner (1860, pp. 71-76) describes his "Methode der eben merklichen Unterschiede (Method of the Just Noticeable

Differences)", which is used in order to measure the so-called "Unterschiedsschwelle (Difference Threshold)". It is interesting that Fechner considers only the procedures of presenting the stimuli in ascending or descending magnitudes without mentioning the possibility of a random presentation of the stimuli, i.e. a randomized presentation as it would be called nowadays. It seems that this holds also for the psychophysicists Delezenne and E. H. Weber who used the "Method of Just Noticeable Differences" even before Fechner as he himself tells us.

In the paper by Peirce and Jastrow (1885), which was read at October 17, 1884, we find (pp. 79-80):

> "A pack of 25 cards were taken, 12 red and 13 black, or *vice versa*, so that in the 50 experiments made at one sitting with a given differential weight, 25 red and 25 black cards should be used. These cards were cut exactly square and their corners were distinguished by holes punched in them so as to indicate the scale of numbers (0, 1, 2, 3) used to designate the degree of confidence of the judgment. The backs of these cards were distinguished from their faces. They were, in fact, made of ordinary playing-cards. At the beginning of a set of 25, the pack was well shuffled, and, the operator and subject having taken their places, the operator was governed by the color of the successive cards in choosing whether he should first diminish the weight and then increase it, or *vice versa*. If the weight was to be first increased and then diminished the operator brought the pressure exerted by the kilogram alone upon the finger of the subject by means of the lever and cam mentioned above, and when the subject said "change" he gently lowered the differential weight, resting in the small pan, upon the pan of the balance. The subject, having appreciated the sensation, again said "change", whereupon the operator removed the differential weight. If, on the other hand, the color of the card directed the weight to be first diminished and then increased, the operator had the differential weight already on the pan of the balance before the pressure was brought to bear on the finger, and made the reverse changes at the command of the subject. The subject then stated his judgment and also his degree of confidence, whereupon the total pressure was at once removed by the cam, and the card that had been used to direct the change was placed face down or face up according as the answer was right or wrong, and with corner indicating the degree of confidence in a determinate position. By means of these trifling devices the important object of rapidity was secured, and any possible psychological guessing of what change the operator was likely to select was avoided. A slight disadvantage in this mode of procedure arises from the long runs of one particular kind of change, which would occasionally be produced by chance and would tend to confuse the mind of the subject. But it seems clear that this disadvantage was less than that which would have been occasioned by his knowing that there would be no such long runs if any means had been taken to prevent them."

This statement shows that the threshold was determined by means of single-case studies, where the sequence of the conditions was fixed by an additional true random experiment in order to avoid any systematic guessing by the subjects. The additional introduction of such a random experiment, in order to control those extraneous variables which are difficult or even impossible to measure, was named randomization by R. A. Fisher (1925).

As for the article by Peirce and Jastrow (1885) note the following:

First, only two subjects participated in the whole study, these being the two authors themselves. I.e., they took it in turns to serve as subject or experimenter.

Second, the article ends with the following interesting conclusion:

> "The general fact has highly important practical bearings, since it gives new reason for believing that we gather what is passing in one another's minds in large measure from sensations so faint that we are not fairly aware of having them, and can give no account of how we reach our conclusions about such matters. The insight of females as well as certain "telepathic" phenomena may be explained in this way. Such faint sensations ought to be fully studied by the psychologist and assiduously cultivated by every man."

This conclusion is of direct importance for an explanation of the surprising performance of "Clever Hans" and other animals as reported in Section 3.3.7.

Obviously, the proposal of Peirce and Jastrow (1885) found no attention by other researchers. Even in the third, revised edition of the textbook "Grundriß der Psychophysik (Foundations of Psychophysics)" by G. F. Lipps (1921, pp. 91-92) merely a word-for-word citation from Fechner's book (1860, pp. 71-72) is given. In view of this it is noteworthy that Dr. T. Erismann, Privatdozent (i.e. a lecturer but still not a professor) at the University of Bonn published a three-volume opus on Psychology at the same time. In the second volume ("Die allgemeinsten Eigenschaften der Psyche (The Most General Properties of Psyche)") he discussed the measurement of sensitiveness, i.e. methods by which the very stimulus may be found which is just noticeable. Here, we find (Erismann, 1921, pp. 93-94, spacing by Erismann):

> "Statt nun das „wissentliche Verfahren" unter verschiedenen Bedingungen wiederholt anzuwenden und dadurch die Fehler zu kompensieren, können wir darauf ausgehen, im „unwissentlichen Verfahren" nach Möglichkeit einen von diesen Fehlerquellen reinen Fall zu schaffen. Wir sagen also der Vp. gar nicht im voraus, in welcher Aufeinanderfolge ihr die Reize dargeboten werden, hüten uns aber zugleich davor, eine bestimmte eindeutige Richtung in der Reizänderung beizubehalten: nach einigen Versuchen könnte sonst die Vp. unser Vorgehen entdeckt haben, und wir würden uns wieder in einem (durch die Unsicherheit des Zeitpunktes, an dem die Vp. ihre Entdeckung macht, noch verschlechterten) wissentlichen Verfahren befinden. Wir bieten ihr also die zwischen bestimmten Grenzen liegenden Werte (im oberen Beispiel wären es also die Werte 2 bis 22 cm) ganz unregelmäßig dar, indem wir nur darauf sehen, daß in einer Versuchsreihe jeder Wert (2, 4, 6 usw.) nicht mehr und nicht weniger als einmal vorkommt. Am besten ist hierbei folgendes Verfahren: man schreibt sich alle Einzelwerte auf Zettel auf, die man in eine Urne hineinlegt; darauf entnimmt man der Urne einen Zettel nach dem anderen, bietet den darauf stehenden Wert und legt den Zettel beiseite. Durch dieses Verfahren ist zugleich alle ungewollte Regelmäßigkeit von seiten des Experimentators ausgeschlossen, wie sie sich in unwillkürlichen Systembildungen, z. B. 2, 8, 14, 20, 4, 10, 16, 22 oder dgl., die bei häufiger Wiederholung ebenfalls von der Vp. erraten werden können, findet. Unter diesen Bedingungen kann also die Suggestibilität der Vp. nicht mehr in Frage kommen, wir haben einen reineren Fall vor uns. Doch hat auch diese Methode, namentlich bei ihrer Anwendung auf die Bestimmung der UE ihre Nachteile, da die an die Vp. herantretende Aufgabe unter diesen „unwissentlichen" Bedingungen schwieriger und komplizierter, und die Einstellung der Vp. für den Experimentator viel unübersichtlicher wird als bei dem wissentlichen Verfahren."
>
> (Abbrevations used: Vp. = "Versuchsperson", UE = "Unterschiedsempfindlichkeit")

In our translation this reads:

> "Instead of using the "knowing procedure" repeatedly under different conditions and thereby compensating the errors, we can assume that we get with the "unknowing procedure" a case which is free of the sources of error. This means that we do not tell the subject in which particular sequence the stimuli are presented, at the same time avoiding a fixed direction of stimulus change: otherwise the subject might see through our scheme after some trials and we would be again in the knowing procedure (though this would be affected by the uncertainty of the point of time where the subject first sees through our scheme). Therefore, we present values between certain limits (in the example above this would be the values 2 to 22 cm) to the subject by making sure that in each trial each value (2, 4, 6 etc.) occurs not more or less than once. The optimal procedure is the following: all values are written on slips of paper which are put inside an urn. Then, one slip after the other is taken from the urn, the respective value is presented, and the slip is put aside. By this procedure each kind of unintentional regularity on the part of the experimenter is avoided, such as it may be found in systematically generated sequences, like e.g. 2, 8, 14, 20, 4, 10, 16, 22. A subject may see through the underlying system of such sequences, if the sequences are repeated often enough. Using the present procedure, any suggestibility of the subject cannot have caused the outcome and we have a purer case. This procedure, however, has also some disadvantages, in particular with regard to the determination of the sensitivity for differences. On the one hand, the task is more difficult and complicated for the subject under "unknowing" conditions, on the other hand, the attitude of the subject is much more difficult to be judged by the experimenter than for the "knowing" procedure."

Though Erismann, just as Peirce and Jastrow, does not use the term "randomization", the proposed procedure corresponds to a carefully planned randomization with the aim to control extraneous variables which cannot be easily measured. As Erismann does not give a reference for the proposed procedure it is not clear whether he knew the article by Peirce and Jastrow (1885) or whether he detected the technique of randomization anew, independent of these authors. It is a fact, however, that Erismann was a Privatdozent at the University of Bonn and that, up to the present day, the library of this university holds the journal "Memoirs of the National Academy of Science", in which the article by Peirce and Jastrow appeared.

1.7 Sir Ronald Fisher and Randomization

The statistician Sir Ronald Aylmer Fisher (1890-1962) was presumably the first to actually use the term "randomization". The additional random experiment conducted in order to control, in an experimental situation, those extraneous variables, which can only be measured with difficulty or not at all, is described in Chapter VIII (Principles of Experimentation) of the first edition of his book "Statistical Methods for Research Workers", which was published in 1925. This chapter was not included in later editions but was extended to a book, the opus "Design of Experiments" which was published for the first time in 1935. This new principle is explicitly depicted in the Sections 5 (Statement of the Experiment), 9 (Randomisation; The Physical Basis of the Validity of the Test), 10 (The Effectiveness of Randomisation), 20 (Validity and Randomisation), 22 (Description of the Experiment), 26 (Validity of the Estimation of

Error), 27 (Bias of Systematic Arrangements), 28 (Partial Elimination of Error), and 31 (Randomisation Subject to Double Restriction) of the book. Fisher (1935) introduces the term "randomisation" (using the British English spelling) by means of a fictitious experiment (Fisher, 1966, p.11):

A lady claims she can decide, judging by the taste of the tea, whether the milk (method M) or the tea (method T) had been first to be poured into a cup. To test this, Fisher proposes the following experiment: eight cups of tea are being prepared, four cups by method M and four cups by method T. Then the cups are given to the lady in a random order and she has to decide for each cup by tasting the tea, whether method M or method T has been used.

One way to generate a random order of the eight cups might be to form first a row of the eight cups in an arbitrary order. Then the numbers 1 to 8 are written on eight cards. The cards are shuffled and one after the other is drawn and successively assigned to the cups in the row. At last, the eight cups are reordered according to the eight numbers from 1 to 8. Alternatively, one might use a table of **random numbers** and mark eight random numbers with a pencil with closed eyes. The first number is assigned to the first cup in the row, the second number to the second cup, etc. At last, the cups are reordered according to the size of the eight random numbers from the smallest to the largest number.

The fact that the cups are presented in a random order is called **randomization** by Fisher. According to him, it is the randomization that guarantees that two groups or conditions will only differ in the values of the causal variables of interest, here the two orders tea-milk or milk-tea, and not in addition in the levels of other possible causal variables (Fisher, 1966, p.18).

Fisher points out that it is not enough to require that all cups of tea are identical in all respects. If this was really possible, we could use a design for which, e.g., the systematic alternating order TMTMTMTM would be chosen. However, the condition of complete equality can never be exactly realized, in any experimental situation. There will always be, though possibly very small, differences between the cups of tea, e.g. in weight, form, surface, absolute and relative quantities of milk and tea, in the time used for pouring-in the two components milk and tea, the consistency and temperature of the tea, etc. In principle we must assume for each experiment that there are always infinitely many causes which cannot be controlled and may yield differences in the outcomes for the different experimental conditions. This can influence the interpretation of the outcome of the experiment.

In the article by Fisher (1926), other techniques of experimental design, apart from randomization, are being discussed. These are, e.g., replication, blocking, randomized Latin squares, factorial designs, and randomized factorial designs. **Blocks** are groups of experimental units which resemble each other with respect to particular selected **block variables**, as for instance animals from the same litter. In **Latin squares** the values of two extraneous variables, i.e. variables which may cause effects though they are not studied, are combined systematically. In **factorial designs** this is done for two or more causal variables which are to be studied. Only the combination of randomization with other techniques of experimental design was really new in this article. The other techniques with the exception of factorial designs had already been discussed in Fisher (1925).

While nowadays the simple but indispensable principle of randomization is generally accepted and forms the basis of each rational experimental design, after Fisher's article (1926) there was a long controversy between R. A. Fisher and F. Yates on the one side and Student (i.e. W. D. Gosset), J. Neyman and E. S. Pearson on the other side. The latter three statisticians did not disapprove of randomization in general. They rather presumed that in certain situations systematic designs ("balanced arrangements") might yield more suggestive outcomes than randomized designs ("random arrangements"). With respect to the fictitious experiment, where first tea (T) or first milk (M) was poured into a cup, Fisher's opponents would have accepted rather a systematic plan of the form TMMTTMMT for the order of the eight cups than an order of the form TTTTMMMM as a result of a randomization. This controversy seems to have ended with an article by Yates (1938), i.e. twelve years after Fisher's article.

SUMMARY

1. Some principles of experimental design are hundreds of years old and are still used today.

2. In the Book Daniel in the Old Testament (167-164 B. C.) the use of a control group is described.

3. A study using a control group as well as replication is described by Athenaeus (between 200 and 300).

4. Avicenna (980-1037) set up the following seven rules for the conduction of an experiment:

   a. Avoidance of more than one simultaneously effective causal variable with respect to the drugs,
   b. avoidance of more than one simultaneously effective causal variable with respect to the patients,
   c. use of control groups,
   d. observation of the change of effect variables after a change in causal variables,
   e. identification of possible extraneous variables,
   f. replication of studies to avoid the interpretation of pseudo-effects which are due to chance,
   g. study of the effectiveness of drugs for human beings and not only for animals.

5. John Stuart Mill (1806-1873) described the difference between experiment and observational study and discussed five methods by which causal conclusions from the outcomes of experiments should be possible:

   a. Method of Agreement,
   b. Method of Difference,
   c. Indirect Method of Difference,
   d. Method of Residues,
   e. Method of Concomitant Variations.

6. Wilhelm Max Wundt (1832-1920) described the difference between experiment and observational study in psychological research. For the experiment the following four principles can be formulated:

   a. Manipulation,
   b. Arbitrariness,
   c. Reproducibility,
   d. Isolation.

7. Randomization is one of the most important techniques of experimental research, since it allows to control not only known but also unknown extraneous variables. This is done by introducing an additional random experiment, a technique, which was already described by Peirce and Jastrow (1885) and rediscovered later by Erismann (1921) and Fisher (1925).

Questions

1.1. What are the two most important techniques in experimental design?
1.2. What can be objected to the design of the prophet Daniel?
1.3. How could the lemon experiment of the Egyptian judge be improved?
1.4. What can be objected to the interpretation given to the outcome of the lemon experiment?
1.5. Give another example for rule 5 of Avicenna.
1.6. Generalize rule 7 of Avicenna.
1.7. Give an example to illustrate the problem which is referred to in the footnote from the book of Mill cited in Section 1.4.
1.8. Give an example for Mill's experimental methods.
1.9. What does Wundt mean by "arbitrary" and what does he not mean?
1.10. Give an example for Wundt's principle of isolation.
1.11. What is gained by randomization in the studies described by Peirce and Jastrow and by Erismann?
1.12. Try to define the term "randomization".
1.13. Name three practical disadvantages of randomization.

2 The Object of Experimental Design

2.1 Dependent and Independent Variables

Those who want to perform an experiment or a study expect to obtain results yielding answers to the questions they have formulated in advance. In behavioral research, in general, the determinants for a given behavior are sought, e.g. the necessary conditions for a child to be successful at school.

To answer this kind of questions, two different types of variables have to be considered: **causal variables** and **effect variables**. Causal variables are also called **independent variables** (abbreviated to **IV**) and effect variables, **dependent variables** (abbreviated to **DV**).

Effect variables or dependent variables are those variables which are measured at the subjects. In general, these are **response variables**, e.g., state of health or the number of correct items in a questionnaire. Sometimes, also characteristics of the subjects, e.g., gender or blood group, are considered as dependent variables. It is a question of definition whether such characteristics can also be regarded as effect or response variables.

Usually, only a finite number of values of an independent variable is chosen, the so-called **levels**. If several independent variables are being considered simultaneously, the corresponding combinations of levels are considered instead of the single levels.

The levels of an independent variable might be ordered in some way, though this does not have to be the case. Thus if the levels, e.g., correspond to different drugs, they generally do not have a natural ordering. If the levels, however, correspond to different doses of a drug, a natural ordering of the levels exists.

The independent variables are also called **factors**. Hence we have **factor levels** and combinations of factor levels.

2.2 Selection of Factor Levels

Quite often the appropriate selection of factor levels turns out to be a crucial problem, as the respective selection might determine whether an existing **causal relation** is detected by the study or not. In the case of a continuous independent variable, e.g. the dose of a drug, it is possible to randomly select a given number of levels from the effective range of the drug. The effective range is defined by all possible doses between the dose corresponding to no effect up to a maximum dose which is neither lethal nor causes any serious **adverse effects**. If the levels of a factor are selected by this kind of procedure the factor is called a **random factor**.

This kind of procedure may also be used for discrete independent variables with a large number of discrete levels. However, random factors are not typical of behavioral research. The experimenter rather tries to select the factor levels according to more or less rational criteria, thus obtaining so-called **fixed factors**. Here, the researcher considers, in general, the two extreme points of the effective range, e.g., the dose zero and an appropriate maximum dose. In addition to that, he or she selects doses which cover, more or less perfectly, the whole effective range. Note for this special case, that drugs in most cases exhibit a logarithmic **dose-response curve**, i.e. the dependent variable is a logarithmic function of the independent variable. In such cases a uniform covering of the effective range is sought, e.g., by doubling the preceding dose in each step, except for the dose following the dose zero. An example for such a uniform covering of an effective range from 0 mg to 32 mg would be given by selecting the doses 0 mg, 1 mg, 2 mg, 4 mg, 8 mg, 16 mg, and 32 mg. One should check for each study individually, if the common practice to relate the dose to the body weight (e.g. 2 mg/kg body weight instead of 2 mg) is really justified for the respective case.

Factor levels should always be chosen such that an apparent difference in the dependent variable can be expected for any two adjacent levels, in order to keep costs low. If foreknowledge exists as to where an effect can be expected, more factor levels should be placed into this region and less factor levels into other regions. If dose-response curves are of interest, i.e. if the functional relationship between an independent and a dependent variable is being investigated, there should be more levels in those regions where maxima or minima of the curve can be expected.

In certain situations where the researcher does not assign the levels to the subjects but uses preexisting values of variables as factor levels (e.g. age, gender, presence or absence of a specific disease), we have a **selection of the levels of the independent variable**. Any putative causal relation which is detected in such a case may not exist in reality and may be a pure **artifact**, i.e. a non-existing pseudo-effect.

2.3 Causal Relations and Intervening Variables

The question of the existence of a **causal relation**, i.e. whether the presumed cause and effect are truly related, is often rephrased to the question of, whether a systematic variation of the independent variable causes a systematic change in the dependent variable; in other words, whether the two variables covary.

Such a **covariation,** of course, does not have to exist since there is no reason to assume that the independent variable can indeed be considered as a causal variable influencing the dependent variable. The purpose of an experiment is to test whether there really exists a causal relation between an independent variable and a dependent variable.

Here, a particular difficulty arises in the behavioral and medical sciences, because in most cases an expected causal relation is not observed for all subjects but only for a majority of the considered subjects. The fact that the causal relation is not found for a minority of subjects is attributed to the effect of a so-called **measurement error,** i.e. to the effect of certain uncontrolled **extraneous variables**. However, another explanation could be that the causal relation is only valid for a subpopulation of subjects which itself is characterized by certain levels of certain subject-related variables, e.g., gender, age, blood group etc.

Some researchers might not accept our talking about causality instead of association, prediction or relatedness. However, our argumentation is very similar to that of many other authors, e.g. Holland (1986) who considered Rubin's model for causal inference (Rubin, 1974, 1978, 1990) among other causal frameworks. Reiter (2000) gives a short introduction into the art of designing causal studies.

Even if clear evidence for a causal relation in a well-planned study is found this does not necessarily imply that the causal structure of the respective variables has been revealed. Consider the following simple example: a subject enters a room and observes that an electric floor lamp lights up as soon as he or she treads on a certain spot on the floor. When the subject steps on this spot again the lamp goes out. This action is repeated ten times and each time the same event is triggered. Thus the subject concludes, having good reasons for this, that there is a causal relation between his or her behavior and the lighting up and going out of the lamp. In behavioral research an analogous result in other circumstances might possibly lead to an article about the detection of a new interesting effect (e.g. "The influence of radio transmissions on the aggressiveness of honey-bees"). In the above example, however, the assumption of a direct causal relationship is not very plausible. A possible explanation of the observed phenomenon might be that a hidden observer is watching the subject's behavior and manipulates the lamp according to his or her behavior. Such an explanation, which some people might find paranoid, might be given by a subject who has watched television shows with a hidden camera or who has read about "typical" psychological experiments. However, a subject with some knowledge in physics might argue that there might be a loose connection in the circuit or in the lamp itself, leading to the phenomenon. In any case nobody will be content with the spurious causal relation in this example, but will look for so-called **intervening variables** between cause and effect.

Obviously, such intervening variables are independent variables themselves, with respect to the dependent variable. Furthermore, it is possible that the observed relation between an independent and a dependent variable is the result of a long chain of intervening variables. One might, for instance, observe that the fruit-crop decreases, if cat-food is expensive, and that it increases, if cat-food is cheap. Consider the following **causal chain**: higher prices for cat-food have the effect that cats without an owner are no longer fed. As a result, these cats catch more birds with the consequence of an increase of insect ravage and a decrease in fruit-crop. The revelation of such causal chains requires, in general, an expensive and extensive investigation which, however, is indispensable, if a real insight into the underlying mechanisms is sought.

Considering intervening variables seems to be appropriate only, if they can be directly measured. Often, intervening variables are only **constructs**, i.e. concocted artificial variables, to which certain properties are ascribed and which allow to "explain" why an independent variable has an influence on a dependent variable. If appropriate **operationalizations** of these constructs, i.e. equivalent variables which, however, can be measured directly, cannot be given, they cannot serve as an explanation for any causal relationship. They rather have the effect of obscuring the true relationship. In the above lamp example we might define a construct by assuming that certain boards of the ground possess a "light drive" which has the effect that a floor lamp goes on or out, if one steps on a certain board. This might be a very simple explanation for the observed phenomenon which, however, does not seem plausible to most people. Nevertheless, in behavioral research similar pseudo-explanations such as "drives" and the like, without any appropriate operationalization are frequently given.

2.4 Ockham's Razor

When examining causal relations one often requires the **Principle of Parsimony** to hold, i.e. that the simplest explanation is applied, if there is a choice of several alternatives. This principle is often called **Ockham's razor** or **Occam's razor,** after the English theologian and philosopher William of Ockham (1285?-1349?) who was born in Ockham in the county of Surrey. It is not quite clear why this principle is ascribed to Ockham, since, as Marilyn McCord Adams (1987, p. 157) points out, the same principle had already been formulated in several versions by Aristotle. William of Ockham, according to the above source, is said to have formulated various different versions of the principle and the original Latin versions with the corresponding references are given (pp. 156-157). In the sequel we give a sort of translation hoping that the meaning is as close as possible to that of the original quotations:

1. There is no sense in doing more than what is necessary.
2. If two arguments are enough to prove the truth of a statement we do not need a third argument.
3. If it is not necessary for other reasons [to consider more than one cause], only one cause should be assumed.
4. More than one cause of an effect should only be assumed, if this is justified by reason or experience or by infallible authority.

Usually, Ockham's razor is cited by merely giving the following cryptic statement, but without any exact reference:

> "Entia praeter necessitatem non esse multiplicanda." (One should not multiply values if it is not necessary.)

The German philosopher Immanuel Kant (1724-1804) includes the above quotation in his "Critik der reinen Vernunft (Critique of Pure Reason)" (1781, p. 652) without reference to William of Ockham as a well-known "Schulregel (rule of school)" of philosophers translating it as "daß man die Anfänge (Principien) nicht ohne Noth vervielfältigen müsse (that one should not double the beginnings (principles) without necessity)".

The Principle of Parsimony should not be interpreted in the sense that the simplest explanation for a detected causal relation is found by just assigning the name of a construct such as "light drive" to the phenomenon. It should rather be regarded as a recommendation to reduce the observed relation to other known simple relations which can be tested by experiments.

At this point, another principle, the **Principle of Testability,** has to be considered, accepting only those explanations which can be tested empirically.

By the way, Kant (1781, p. 656) also formulated a kind of counterprinciple to the Principle of Parsimony:

> "Entium varietates non temere esse minuendas."

This means that it is not always sensible to reduce the explanation of a relation to a single cause, since reality is rather complex. Therefore, the Principle of Parsimony should not lead to artificial "simple" explanations despite a complex reality.

2.5 Constructs

The notion of a **construct** is not only used with regard to intervening variables, but also when dealing with dependent and independent variables. If the influence of the intelligence of children on their learning ability is examined, the independent variable "intelligence" as well as the dependent variable "learning ability" are constructs which cannot be measured directly, but have to be operationalized first. For example, one might operationalize "intelligence" as the score in an intelligence test and "learning ability" as the number of correctly solved problems after a preceding training. Another name for construct is **latent variable**, while the respective operationalization of a construct is denoted as the corresponding **manifest variable**. Different researchers may use different operationalizations of the same construct, e.g. different intelligence tests, different methods of training, different problems etc., and might, as a result, draw different conclusions about causal relations between the same constructs.

2.6 Causal and Correlative Relations

If one is interested in the relation between two variables, regarding one variable as the independent and the other one as the dependent variable does not always make sense. Possibly none of the two variables causally influences the other variable. For example, if we measure children's ability to calculate and to spell, we can try to enhance the ability to calculate by a systematic training, which does not imply that the ability to spell is also improved. Similarly, a training in spelling will raise the spelling ability, but not necessarily the ability to calculate at the same time. A high correlative relation between the two variables, however, which has existed even before the training has not been ruled out, i.e. there may be a high percentage of children scoring high in both variables while another considerable percentage of the children only scores low.

First one can, thus, conclude that high **correlative relations** between variables might be observed though corresponding **causal relations** do not exist. Second, such a correlative relation should make the researcher search for the cause of such a relation, i.e. for an independent variable which influences the two dependent variables "ability to calculate" and "spelling ability" (in our example in the same direction). An obvious candidate for such a causal variable could be, e.g., the number of hours a child gets support in these fields by the parents and other persons outside school. This number might be taken as an operationalization of the construct "amount of care". Another causal variable might be "intelligence". This construct, however, is probably more difficult to operationalize than the construct "amount of care".

In any case, such a **third variable** which leads to the **illusory correlation** between our first two variables, has the property, that for fixed values of the third variable a correlation will not be observed between the two first variables. For instance, consider "amount of care" as such a third variable. The abilities to calculate and to spell should not correlate at all, if children receiving the same amount of care are being considered.

SUMMARY

1. Experiments are performed in order to prove the existence of causal relations between cause and effect variables.

2. The researcher can either fix the values of the cause variables systematically or select them randomly.

3. Observed causal relations do not necessarily have to be an evidence for a direct causal relationship, but might also be traced back to the effects of a chain of intervening variables.

4. According to the Principle of Parsimony (Ockham's razor) a simple explanation should always be considered.

5. Constructs or latent variables are non-observable variables for which corresponding observable manifest variables have to be defined.

6. If high correlations are being observed, i.e., if changes in effect variables parallel corresponding changes in cause variables, nevertheless, no guarantee for the actual existence of a causal relationship is given.

Questions

2.1. What are other names for DV and IV?

2.2. Is it possible that dependent variables are latent variables?

2.3. Which rules should be obeyed when the levels of an independent variable are being determined?

2.4. Why is it difficult to draw causal conclusions if the levels of an independent variable are the result of selection?

2.5. Why is it not possible to conclude a causal relation from a correlative relation?

2.6. Explain the terms "intervening variables" and "causal chains".

2.7. Suggest an operationalization for "nervousness".

2.8. What do we understand by Ockham's razor?

2.9. In a study (cf. [4]) the results for five samples with a total of 663 subjects in a Psychosomatic Attitude Questionnaire (PEF), a humor test, as well as several measures of conservatism were obtained. The correlations between the ten subscales of the PEF, the measures of conservatism and the rating of the funniness of incongruity humor were calculated. A conclusion was that the acceptance of incongruity humor is related to conservatism.

Discuss why such a study is not suited for getting information about the real relation between the constructs "incongruity humor" and "conservatism".

3 A Case for Experimental Design

As mentioned above, experimental design is crucial for a causal relationship between an independent and a dependent variable to be proved, in as far as that a discovered relation cannot be conclusively interpreted as a causal relation without the appropriate experimental design. It is possible, e.g., that not the considered independent variable but a totally different causal variable which has not been considered, is responsible for variations of the dependent variable. Of course, such a variable will be connected with the respective independent variable in some way, though we probably do not understand this connection. In our example in Section 2.3 where we were able to switch on and switch off a floor lamp by stepping on a board, such a causal variable, which we actually did not consider, may have been a person (e.g. a psychologist) observing us, though we were not aware of this, and switching the lamp on and off, depending on our behavior. If we eliminate the true cause, e.g. by removing the hidden observer, the putative causal relation disappears. I.e., a real causal variable is confounded with the independent variable and this **confounding** yields a spurious, non-existing causal relation.

A similar problem occurred when we explained intervening variables. If there are one or even more intervening variables between the independent and the dependent variable, we may regard intervening variables as confounding variables, if we cannot eliminate them without influencing the spurious causal relation between independent and dependent variable. However, according to the definition of intervening variables this kind of influence should always exist.

A confounding variable cannot only trigger a non-existing causal relation but also conceal an actually existing relation. In our example the secret observer might always compensate our actions, thus preventing any potential effect of our steps on the board on the lamp, though there truly exists a corresponding physical relation. Here one cannot detect the existing causal relation unless the confounding variable, i.e. the secret observer, is eliminated.

The objective of experimental design is to render each kind of **alternative explanation** for a detected causal relation implausible. One aims at making the validity of a causal conclusion unassailable to any possible objection. By means of experimental design, the researcher tries to provide arguments against all potential **threats to validity**, i.e. against any alternative explanations, on an empirical basis.

It is obvious that it is impossible to avoid all these threats to validity, if we admit the existence of thought-reading rats or invisible demons who intervene in our experiments. It is, thus, only necessary to rule out those alternative explanations whose validity can be tested. Depending on the individual problem and design, however, a whole set of these alternative explanations might exist. One of the most comprehensive compilations of possible alternative explanations, which is also the basis of the present chapter, can be found in the important opus "Quasi-Experimentation" by Cook and Campbell (1979), who treat the set of typical alternative explanations under the following headings:

1. Statistical Conclusion Validity,
2. Internal Validity,
3. Construct Validity of Putative Causes and Effects,
4. External Validity.

In fact, the first point is regarded as being part of the second, and the third as part of the fourth. The authors denote the validity of the causal conclusion from the independent to the dependent variable as **internal validity** (Campbell, 1957). Here, **statistical conclusion validity** has to be considered since, as a rule, only a very small part of the extraneous variables can be controlled or is known in medical or behavioral research. Therefore, we assume a valid causal conclusion even if the causal relation cannot be detected in a small percentage of cases as long as it is found in a sufficiently high percentage of replications.

The term **external validity** (Campbell, 1957) or **generalizability** is used to describe the situation where a found causal relation is still valid, even if the conditions are far less restricted than those in the experiment, where the existence of the relation was proved. In particular, **construct validity** of putative causes and effects refers to the validity of a causal relation for the constructs themselves, whose existence was proved in the experiment only for the respective operationalizations of the constructs.

Some researchers, however, use the notions internal and external validity to legitimate meaningless pseudo-research. They claim, and this is not correct, that there is some kind of equivalence between these two types of validity and declare that, in their field of research, external validity is of greater importance than internal validity. By doing so, they try to justify studies in which conclusions are drawn and the existence of effects is claimed, though these declarations do not have any value due to many plausible alternative explanations, even if such pseudo-results were obtained in a very general situation. Generalizability of results lacking any foundation is meaningless from a scientific point of view. There is, beyond dispute, a priority of internal validity over external validity: external validity without sufficient internal validity is meaningless.

Nevertheless, studies with even a high internal validity might not be important at all, e.g. if the examined situation is so restricted that only the existence of trivial or unimportant relations or effects has been proved. However, the situation in such cases tends to be so obvious that hardly anyone will be inclined to overestimate such results.

The organization of the following explanation largely corresponds to that in Chapter 2 in Cook and Campbell (1979).

3.1 Threats to Statistical Conclusion Validity

For the following Sections 3.1.1, 3.1.2, and 3.1.3 we assumed the reader to be informed about the fundamentals of statistical tests. If this is not the case these sections may be ignored without any harm.

3.1.1 Low Statistical Power

Power is defined as the probability to detect an effect by means of a statistical test if this effect exists in reality. One can assume for all reasonable statistical tests that the power is the higher the larger the sample size, the larger the size of the effect, and the smaller the variance of the dependent variable. From this one can conclude that a statistical test has a low power, i.e. the probability of detecting an actually existing effect is low, if the sample size is small or if the size of the effect is small or if the variance of the dependent variable is large.

If the sample size is small in relation to the variance of the dependent variable and if the size of the effect of interest is small it is possible that one gets a **null result** i.e. one can neither prove the existence of an effect nor is it allowed to claim that no effect exists. In such a case the outcome of the experiment does not allow any conclusions. This impossibility to prove the non-existence of an effect or a causal relation is due to the fact that we can never rule out the possibility that we might prove the existence of an effect if only the sample size was large enough.

Though we know that this cannot really be done one sometimes intends to prove the equivalence of two treatments, i.e. the so-called bioequivalence problem is being considered. As it is not possible, as shown above, to prove the exact equivalence one tries to show, that the effects of two treatments of interest do not differ from each other by more than a given amount which is unequal to zero. But even for bioequivalence tests null results for which no conclusion is possible can occur.

3.1.2 Violated Assumptions of Statistical Tests

In general, statistical procedures yield valid results, only if certain assumptions hold. The most important among these assumptions seem to be the independence of the measurements, the homogeneity of the distributions of the measurements for the subjects within a sample, the normal distribution of the measurements, the equality of the variances of the measurements for different samples and in certain situations linear relationships between independent and dependent variables. The independence of the measurements is assumed if the subjects in a study cannot communicate with each other in the course of the study and if only one measurement is considered for each subject. The independence of the measurements is not guaranteed, e.g., in the case of **repeated-measures designs**, where additional assumptions concerning the dependence structure of the measurements have to be fulfilled. The homogeneity of the distributions is assumed if all subjects are randomly assigned to the different experimental conditions (**randomization**). The assumption of normal distributions can never really be justified. The same holds for the assumption of equal variances because it is not plausible that only the means but not the variances should be affected by different experimental conditions. Finally, it is impossible that a linear relationship exists in reality between an independent and a dependent variable, though an approximate linear relation may be observed if the values of the independent variable are neither too small nor too large.

In particular, with respect to such assumptions as normality or equality of variances, one never knows, when dealing with real data, if these assumptions are violated or not. Hence, one has to be also careful with regard to this problem when interpreting the outcome of an experiment with respect to the existence of a causal relation. By the way, statistical procedures used to test the validity of such assumptions are not of much value because one can at most conclude that the assumptions do not hold. As a rule, one ought to dispense with all parametric significance tests, i.e. with those tests, which presume the variables to be normally distributed, an assumption one can hardly ever justify. Thus, one had better use distribution-free or robust procedures which require much weaker assumptions. A new generation of distribution-free statistical tests for which not only the assumption of normality is unnecessary but which also permits dependent observations and unequal variances is of particular interest (Brunner and Denker, 1994; Brunner, Puri and Sun, 1995).

3.1.3 Multiple Tests

If the decision that a particular effect exists is taken on the basis of data, this decision might be wrong, because, e.g., random fluctuations of the data might have given the wrong impression that an effect is present though in reality it is not. A statistical test takes this possibility into account by assuring that the probability of a wrong decision of this kind is at most equal to a given level α (alpha). Usually, this level is fixed to be $\alpha = .05$. In other words, if a statistical test yields a significant result at the level α this does not mean that the corresponding effect exists in reality. However, the probability that a test result is significant though no effect is present in reality is at most equal to α.

If a sample of subjects is being considered for which several dependent variables are recorded, one could be inclined to test for as many causal relations. However, the probability of detecting non-existing pseudo-effects which are caused by random heterogeneity of the sample is increased with the number of such tests. If there are not too many of these tests, a statistical correction can be performed by means of an alpha-adjustment for multiple tests.

The oldest and best-known procedure of this type is the Bonferroni-correction which at the same time has the weakest assumptions of all comparable alpha-adjustments: if a total of k statistical tests are to be performed at level α, the single tests are performed at level (α / k) as opposed to α. If a significant result is being found for this adjusted level, the result is regarded as being as significant at level α. When using a procedure like this one can be sure that the probability of finding one or more pseudo-effects is not larger than α.

A sensible research strategy requires the number of effects which one wants to test in a study to be fixed in advance. If more than one test is planned, an alpha-adjustment has to be conducted. Data related to those effects, which are less important for the time being and which are not the focus of the present study, will also be evaluated. However, one does not consider whether these results are "significant" but uses them for generating new hypotheses or for supporting the interpretation of the main results. If the possible existence of an interesting effect is indicated due to such a supplementary evaluation, a new independent study with new subjects is necessary for proving this existence.

3.1.4 Reliability of the Dependent Variables

When dependent variables are being measured, systematic as well as unsystematic errors of measurement may be the reason for existing causal relations not being detected or non-existing causal relations being found. A high variance of the measurement error is equivalent to a low reliability of a dependent variable. The effects of both kinds of measurement errors should be kept low by careful selection and examination of the measuring instruments used. I.e., the experimenter ought to avoid spurious effects by a suited experimental design. An existing low precision of measurement which is due to unsystematic errors of measurement otherwise has to be compensated for by large sample sizes (cf. Section 3.1.1).

3.1.5 Reliability of the Independent Variables

If the independent variable is a **quantitative factor**, given, e.g., by the doses of a drug, similar errors of measurement as in Section 3.1.4 can be expected, though, in general, to a less extent. If, however, the independent variable is a **qualitative factor**, e.g. defined by different treatments, the treatments must not be applied differently by different experimenters or by the same experimenter at different points of time or for different subjects. Here, a strict **standardization of treatments** is necessary. A high variance of the measurement errors or a low level of standardization are equivalent with a low reliability of the independent variables.

3.1.6 Random Disturbance of the Experimental Situation

Examples for a random disturbance of the experimental situation are unexpected noise caused by road-construction measures during a laboratory experiment or the unintended dropping of a noise-producing object. Such incidents can increase the error variance and, thereby, become a threat to statistical conclusion validity. This is because a high error variance causes a low probability of detecting effects which exist in reality, i.e. a low power of statistical tests.

3.1.7 Random Differences between Subjects

The more heterogeneous the subjects, the higher is the error variance and the more difficult is a proof of the existence of causal relations. If it is not possible to remove the causes of this heterogeneity, the only way out is the use of a sufficiently large sample size. Often, random differences between subjects are assumed to be caused by uncontrolled **extraneous variables**.

3.2 Threats to Internal Validity

3.2.1 History

If subjects are measured before and after a treatment, where the "treatment" may also be a control condition without the experimenter actually intervening, the term **history** is used in order to describe those non-planned occurrences of any incidents between the two measurements which, in some way, may influence the second measurement. For instance a failure of the heating in a laboratory in the time interval between the two measurements might serve as an example to illustrate this.

3.2.2 Maturation

If a measurement is being conducted before and after a treatment, the fact that the subjects may change in the time interval between the measurements must be taken into account as well as that this change might not be related to the experimental situation. It is obvious, for instance, that if a lot of time has passed between the measurements, the subjects have got older in the meantime and have changed in a number of ways. This will be called **maturation** in the sequel. E.g., in a learning situation we might have measured the velocity of young rats running over a narrow foot-bridge. If we repeat these measurements two months later, they cannot simply be compared to the former measurements, as the animals will be heavier such that their velocity will, thus, be influenced.

3.2.3 Testing

If the same subject is measured repeatedly, a **habituation** to the measurement procedure cannot be ruled out. Apart from this, **sensitization** might occur, where memory effects may be of importance. In such cases, the results of different measurements are difficult to compare. Corresponding misinterpretations are said to be the effect of **testing**. If, e.g., the heart rate is measured it might be that the first measurement will differ from the following measurements because the subject habituates to the measuring procedure. On the other hand, if a subject is asked, after a first treatment, whether he or she experiences a kind of nausea, this subject becomes sensitized and might respond in a different way to the same question after further treatments.

3.2.4 Instrumentation

In the course of a study, some of the characteristics of the measurement might change, if the subjects are measured repeatedly. One reason for this may be, e.g., that the experimenter has gained experience. Furthermore, a measuring device might not have the same precision at all points of a scale. In particular, **ceiling effects** (the precision is low for high scale values) or **floor effects** (the precision is low for low scale values) might occur. The change of a measuring device during a study in any ways is said to be the effect of **instrumentation**.

3.2.5 Statistical Regression

Consider the following situation: A researcher has developed a method to improve the result of learning of children. The construct "learning result" is measured by a scale with a maximal score of 100 points. The researcher presumes that a **ceiling effect** might occur for those children who have already scored high before the application of the new method, as these children cannot improve very much unless a different scale is used. According to the researcher's opinion including these children in the study would result in an underestimation of the method's beneficial effect. Therefore, by means of the scale a screening is performed for a large sample of children, and only children with 10 or less points are included in the study. Consequently, the effectiveness of the method is measured by the increment in points. Obviously, there might be objections against this kind of design. In particular, the increment in points might be ascribed to the alternative explanation of **statistical regression,** which is also called **regression to the mean**.

For our example, this can be explained as follows: The scale by which the children are assessed, underlies unpredictible fluctuations due to **measurement errors**. These are, however, not necessarily errors of a physical device. For example, the assessment scale used here is always the same from a physical point of view. Everything which might influence the performance of the children in the assessment and which is not directly related to the experimental manipulation is regarded as a measurement error. If a child did not sleep well the night before the assessment, it will possibly have a worse result than under different circumstances. By contrast, if the child studied the relevant topics very intensively the day before, it might score higher than it usually would.

Therefore, measurement errors may cause larger as well as smaller scores in comparison with the "normal" or "true values". If we assume that a high percentage of children with small scale values before the treatment got these results because of negative measurement errors, we may expect that for many of these children the second measurement will yield a higher value, irrespective of any intervention, because the probability is low that after a negative measurement error from the first recording an even more negative measurement error will follow during the second recording. Similarly many children with a high initial value will show a lower value at the second measurement. By this effect certain subgroups of children will exhibit a tendency towards the mean value though the variance of the values may be unchanged with respect to the whole sample. As a consequence, the higher second values in our example, after eliminating children with high initial values, need not to have resulted from the method for improving the learning results.

The terms statistical regression or regression to the mean or **regression artifact** are difficult to understand in the context above for someone who is familiar with statistical regression analysis today. The first one to mention this artifact was Galton (1877) who called it *reversion*. Later on it was termed by him *regression towards mediocrity* in Galton (1885). In this latter article we find for the first time figures which show regression lines. All in all, we see that Galton's naming of the observed artifact is understandable, while modern statistical regression analysis has not much to do with the etymological meaning of the term "regression".

Galton believed that reversion or regression toward mediocrity was a real effect and not only an artifact as we illustrated above. About fifty years after Galton's detection of reversion, another researcher found a similar effect which has the same direction as statistical regression. This is the **Law of Initial Values** by Wilder (1931). Wilder and many others believed to have detected a physiological law claiming that after low physiological measurements high ones will follow and vice versa. These researchers are convinced that this effect is not due to statistical regression and, therefore, no statistical artifact (cf., e.g., Berntson et al., 1994). Wilder (1931) applied pilocarpine, atropine, and adrenaline to human subjects and measured pulse and blood pressure. He describes the effect that in all situations in about 75 percent of the trials the increasing curve was the flatter and the decreasing curve was the deeper, the higher the initial values of pulse and blood pressure and vice versa.

3.2.6 Selection

If a group of subjects receives a treatment and another group does not and both groups are drawn from different populations, a difference of the results for the two groups, with respect to the dependent variable, is not necessarily due to the treatment. It may well be that the difference in the results is solely caused by the difference of the populations, while the difference in the experimental conditions (treatment versus control) did not have any effect. Of course, such a **selection effect** can also lead to the result that an actually existing treatment effect is not found or that instead of an existing positive treatment effect a spurious negative treatment effect is observed.

If groups are part of different populations, a selection effect can occur, if the velocity of **maturation** (cf. Section 3.2.2) is different for the populations. Similarly, differences in **history** (cf. Section 3.2.1) for different populations may result in a selection effect. Also **instrumentation** (cf. Section 3.2.4) may cause selection effects, if the different populations differ in the range of values which are attained by the dependent variable.

3.2.7 Experimental Mortality

Sometimes it is not possible to measure the dependent variable for all subjects who originally participated in the study. This might be due to several reasons: A subject might, indeed, have died in the meantime, or it is no longer willing to take part in the study, or it is no longer available, because it moved to another town. All of these cases are called **experimental mortality**. If one cannot rule out that different levels of the independent variable lead to different **dropout rates**, a causal interpretation of the outcome is difficult. The reason for this is that the remaining subjects are a selection of the original sample which might thus lead to a selection effect (cf. Section 3.2.6).

3.2.8 Direction of the Causal Conclusion

In the example in Section 2.6 a relation between the abilities to calculate and to spell was assumed. If the ability to calculate is used as an independent and the ability to spell as a dependent variable, a study might reveal that a high ability to calculate renders a high ability to spell. In another study, however, the roles of independent and dependent variable, respectively, might have been interchanged and thus the opposite causal relation is found, i.e., a high ability to spell triggers a high ability to calculate. If, as assumed in Section 2.6, a **third variable** is found, which is the true causal variable and influences the ability to spell as well as the ability to calculate, e.g. the variables "amount of care" or "intelligence", there is no problem. Nevertheless, if for all presumed third variables which are imagined by the researcher no empirical evidence for an influence on our two abilities is found, an interpretation of the outcome will be very difficult, because two opposing causal relations cannot be valid at the same time.

3.2.9 Exchange of Information

If one cannot avoid that subjects are able to interchange information about the study, this may have the consequence that existing effects may not be found or are artificially augmented or even effects with a direction contrary to reality are found. First, it is possible that groups which originally behave differently become so similar in their behavior that causal conclusions become impossible (**diffusion of treatments, imitation of treatments**). Second, groups which are exposed to a far less attractive condition may try to compensate this disadvantage (**compensatory equalization of treatments, compensatory rivalry**). Such an overcompensation was also called "**John Henry Effect**" (Cook and Campbell, 1979, p. 55). John Henry was a steel worker who knew his output was to be compared with that of a steam drill. He succeeded in outperforming the machine, but consequently died of overexertion. Finally, it is possible that groups with obviously worse conditions than other groups are demotivated with respect to their tasks which may result in an artificial increase of the differences to be expected (**resentful demoralization**).

3.3 Threats to Construct Validity

3.3.1 Inexact Definitions of Constructs

With respect to constructs, causal conclusions can only be drawn from empirical outcomes, if the constructs are defined sufficiently precise, i.e. sufficiently restrictive. E.g., the construct "anxiety" is far too general, if causal relations are to be detected. It is obvious that, e.g., "anxiety caused by examinations" and "anxiety caused by height" are two constructs which have not much in common though both are aspects of "anxiety". This does not change if we restrict the construct "anxiety" to the construct "situation-related anxiety", because "anxiety caused by examinations" as well as "anxiety caused by height" can be subsumed under this label. But even the construct "anxiety caused by examinations" is not precise enough: Which kind of "anxiety" is meant? Is it the anxiety to fail the examination or the anxiety with respect to possible consequences of a failure or is it fear of the examiner? Are there "cognitive" causes of the anxiety or is it a response to physiological reactions caused by stress? Each of these possible constructs, all corresponding to the label "anxiety caused by examinations", which may be defined by these or other considerations must be operationalized separately. The validity of causal relations which are found for such specific constructs must not be generalized without a new empirical test to the other constructs and in particular not to the more general constructs "anxiety caused by examinations" or even "anxiety" without any restriction.

3.3.2 Mono-Operation Bias

If the influence of "noise" on "memory" is being studied, only weak statements can be formulated if only one kind of noise and one aspect of memory is considered. The study should comprise different kinds of noise with respect, e.g., to meaningfulness and loudness, and different aspects of memory, e.g., with respect to the material to be learned and the length of time between learning and recall.

3.3.3 Mono-Method Bias

When the influence of "noise" on "memory" is being studied, noise should not only be given via headphones and memory should not only be judged by a reproduction task. Otherwise one cannot rule out that totally different effects may be observed if no headphones are used or if the saving of time in relearning is used as a measure of memory performance.

3.3.4. Hypothesis Guessing

At least for subjects who know that they participate in a study one has to take into account that they form certain hypotheses about the purpose of the study. Depending on the specific instructions and situations this may have the consequence that the behavior of subjects admittedly allows to draw causal conclusions but, unfortunately, not with respect to those constructs in which the researcher is interested in. An example for this is the so-called **Hawthorne effect** (Robinson, 1976, pp. 97-98). This effect is named after a study in the Hawthorne plant of the Western Electric company where one found out that, regardless of which changes in working conditions were implemented, the production increased.

One possible explanation for this phenomenon is that the working women observed that something was altered in lighting, working hours or other conditions, so that they formed hypotheses about the reasons for these changes, and in consequence altered their working behavior at least for the time of the study. Meaningful causal conclusions would have been only possible if the working women had continued to work with the same motivation as before and had not consciously observed the changes, e.g., in lighting.

3.3.5 Social Desirability Responding

A further threat to construct validity occurs if subjects try to behave in a way that makes them appear competent and "normal" to the experimenters. The behavior shown can be influenced considerably by age, gender, and behavior of the experimenters.

3.3.6 Experimenter Expectancies

Each study is performed to test one or more hypotheses which were formulated in advance. As a consequence, experimenters have certain expectancies with respect to the outcome of a study. Even though they try to be as "objective" as possible in their behavior toward the subjects, an unconscious and unintentional influence can never be ruled out. Studies of this phenomenon were performed by Rosenthal (1966). Thus, this effect is also called **Rosenthal effect**.

In clinical trials **single-blind studies** are performed where the patient does not know which treatment is being given. In **double-blind studies** neither the doctor nor the patient knows which treatment is being applied. These types of studies are thought to be a kind of protection against the occurrence of the Rosenthal effect. Finally, in

triple-blind studies not only the patient and the doctor do not know which treatment is being used, but in addition the person analyzing the data does not know what data belong to which experimental condition.

3.3.7 Clever Hans and his Friends

Actually, the effects of the experimenter's expectancies (Section 3.3.6) on the outcome of experiments, in particular with respect to experiments on animals, have been known for a long time. Here, we would like to recall Wilhelm von Osten who invented the "Klopfsprache der Tiere (knocking language of animals)" and first instructed a horse in Berlin in 1890 and later another horse which is known as the "kluge Hans (Clever Hans)". The animals learned to count, by knocking the numbers with a hoof on the floor and also could solve simple arithmetic problems. Later, the animals learned to spell using a table where each letter was coded by a number. In Elberfeld Karl Kraus instructed two young Arab stallions, called "Muhamed" and "Zarif", as well as several other horses, one of them completely blind. The Arab stallions could not only extract square and cubic roots from powers but could also reply, naturally in German. Similar interesting records about talking dogs and cats (e.g., "Paula Moekel: Mein Hund Rolf, ein rechnender und buchstabierender Airedale-Terrier (My Dog Rolf, a Calculating and Spelling Airedale), Verlag von Robert Lutz, 1919" and a subsequent volume "Erinnerungen und Briefe eines Hundes (Memoirs and Letters of a Dog)") were published at least up to 1920 in several books and in articles in the journal "Mitteilungen der Gesellschaft für Tierpsychologie (Communications of the Society of Animal Psychology)". Professor Dr. Heinrich Ernst Ziegler, who was professor at the Technische Hochschule in Hohenheim, provides us with detailed information on this topic in his book entitled "Tierpsychologie (Animal Psychology)" (1921, pp. 65-73). This book is a shortened version of a lecture on Animal Psychology which Professor Ziegler had been giving since 1890 at the University of Freiburg im Breisgau, at Jena, and at the Technische Hochschule in Stuttgart.

However, all these results were regarded by the scientific community with some scepticism. Dr. Oskar Pfungst, in his book "Das Pferd des Herrn von Osten (The Horse of Herrn von Osten)" which was published in 1907, tried to prove, by means of extensive empirical studies, that "Clever Hans" was able to translate scarcely perceptible body changes of his master in knocking signals (Pfungst 1907/1977). This book was published in English in 1911 by Holt in New York and a reprint edited by R. Rosenthal (cf. Section 3.3.6) was published in 1965 by Holt, Rinehart and Winston in New York.

It is not astonishing that the results of the investigations by Pfungst and others were criticized as erroneous and rebuffed by the faithful such as Ziegler (1921, p. 66). In this context Ziegler (1921, p. 69) refers to his publications on the talents of his dog Awa, which was the son of the talking she-dog Lola which in turn was a daughter of the dog Rolf of Frau Moekel who was mentioned above. The whole discussion reminds us, superficially, of certain publications or media reports on the abilities of autistic children.

Nevertheless, these reports about talking dogs, horses, and cats (as for the cats compare Ziegler, 1921, p. 69, p. 71) should make us draw the conclusion that in any study, either with animals (e.g. rats), or with human beings (e.g. students) the researcher's hypotheses should not have any influence—by non-conscious or unintentional hints—on the behavior of the subjects being studied (cf. Section 4.10.5).

3.3.8 Omitting Relevant Levels of Constructs

If the influence of noise on memory is being investigated, no causal relation will be found if, e.g., only five levels of noise below the threshold of audibility are selected and in each case only one syllable is to be learned. A relation between the two constructs can only be proved to exist if the levels of the constructs are selected in such a way that a distinct relation may be expected. Based on the outcome of a **preliminary experiment** a set of promising levels is tried to be selected for both constructs.

3.3.9 Effects of more than One Independent Variable

In studies where the subjects are exposed to several independent variables it might be difficult to generalize observed effects on the dependent variables. First, it is not known, whether the found effect occurs only in situations where more than one independent variable is active. Second, it is not known, whether the effect of a given independent variable is the same, if this variable is solely active or in context with other variables.

3.3.10 Interaction of Testing and Treatment

One cannot be sure from the start that a causal relation which has been found exists independent of whether one has one or more measurements for each subject (cf. Section 3.2.3). E.g., if the condition of subjects is measured by means of a questionnaire it may be a difference whether the subjects get the questionnaire only as a **posttest** or also as a **pretest**.

E.g., assume that a treatment consists of the application of a drug and that you consider a treatment group of subjects which gets the drug and a control group of subjects which gets a **placebo**, i.e. a bogus drug which does not contain any effective ingredients. Thus, the independent variable has the two levels "drug" and "placebo". The dependent variable "nausea" is measured by asking the subjects whether they experience nausea with the two possible answers "yes" or "no". If we ask only after having given drug or placebo, respectively, we may observe that all subjects in the drug group answer "yes", while all subjects in the placebo group answer "no". Here, we observe a causal relation in that sense that the drug causes nausea or better: that the drug causes people to answer "yes" to the question about nausea. However, if we consider another treatment and control group, where we pose the question before and after the drug or placebo, respectively, has been applied another outcome might result. E.g., it might be observed that all persons in both groups answer with "yes". In this case no causal relation can be detected because no association between drug and nausea is observed. The reason for this outcome might be that the subjects are sensitized by the first question, i.e. they expect that the treatment will cause a nausea and as a consequence they experience a nausea in both levels of the independent variable.

Such an outcome demonstrates an interaction between testing and treatment because in one test situation we find a difference between treatment and control condition, in the other test situation we do not find such a difference. If we had found in the test situation with a pretest the same difference between the two treatment conditions with respect to the dependent variable which we observed in the situation without a pretest, we would not have had an interaction between testing and treatment.

3.3.11 Restricted Generalizability over the Constructs

If a causal relation has been found which describes the effect of different kinds of noise on different kinds of memory performance a conclusion that similar laws are valid for those types of noise and those types of memory performance which have not been considered in the study is not allowed.

3.4 Threats to External Validity

3.4.1 Interaction between Selection and Treatment

In almost every study the subjects are a sample drawn from a not well defined subpopulation. If certain causal relations have been proved it is uncertain whether these relations also hold for other subpopulations for which no studies exist.

E.g., imagine that you select a sample of 40 children between eight and ten years. You randomly split this sample into two subsamples of the same size (20), one of which gets a newly developed memory training, the other does not. Then the children have to remember the names of 100 sportsmen. As a result you find that the group with the memory training shows a far better performance than the other group, i.e. the training has had an effect. Another researcher replicates your experiment and does not find any difference between the two samples, i.e. no effect of the memory training. You now observe that in your initial sample 90% girls were present, while the initial sample of your colleague contained 90% boys. Because the boys knew most of the sportsmen before the experiment, we have a **ceiling effect**, and no effect of the training can be observed. Thus, a causal relation between memory training and memory performance can only be observed in the subpopulation of girls but not in that of the boys. Therefore, a difference between training and no training can be only observed in one subpopulation but not in the other. This means that we have an interaction between treatment and gender. If the outcome would have been the same for both sexes, no interaction between the two populations would have been present. Since gender is only one possible characteristic for describing the result of a selection, we have, more generally, an example for an interaction between selection and treatment in the above experiment.

3.4.2 Interaction between Setting and Treatment

In every study only certain **settings** can be realized. It is uncertain whether causal relations whose existence has been proved for one situation may be transferred to situations which were not studied.

E.g., assume that you select a sample of 80 children between eight and ten years. You randomly split this sample into four subsamples of the same size (20). Two subsamples get a newly developed memory training, the others do not. Then the children have to remember the names of 100 sportsmen. One sample with training and one sample without training are tested in a laboratory room without windows. For these groups a training effect is observed. The other two groups are tested in a classroom, where it is possible to observe through the windows other children playing outside. For these latter groups no training effect is observed. The reason for this is obvious: in the second setting the children are diverted and do not differ in their performance due to a **floor effect**. Thus, we find an interaction between setting and treatment. If the children had not been diverted in the second setting the outcomes for the two settings would have been similar and no interaction between setting and treatment would have been observed.

3.4.3 Interaction between History and Treatment

If a causal relation is found for subjects at a given time, it is uncertain, whether the same relation would have been found ten years ago or whether it will be found in the future, because the general social situation may change or fundamental attitudes may change due to the occurrence of e.g. wars or migration.

E.g., assume that you select a sample of 40 children between eight and ten years. You randomly split this sample into four subsamples of the same size (10). Two subsamples get a newly developed memory training, the others do not. Then the children have to remember the names of 100 sportsmen. One sample with training and one sample without training are tested in a laboratory room on a given day. The other two samples are tested in the same room the next day. You find an effect of the training for the first day but not for the second day. A possible reason for the latter outcome is that in the night between the two days a fire has devastated a considerable part of the town and most children have not slept much in that night. As a consequence, both groups exhibit a low performance and do not differ due to a **floor effect**. Thus, we have an interaction between history and treatment. Without the fire a similar outcome would have been observed for both days and no interaction between history and treatment would have occurred.

SUMMARY

1. Experimental design is necessary, because otherwise alternative explanations (threats to validity) for observed effects cannot be ruled out.

2. Four kinds of validity have been considered:
   a. Statistical conclusion validity,
   b. Internal validity,
   c. Construct validity,
   d. External validity.

3. Threats to statistical conclusion validity occur amongst others in the context of low statistical power, violated assumptions of statistical tests, multiple tests, low reliability of dependent and independent variables, random errors and random differences of subjects.

4. Threats to internal validity occur amongst others in the context of history, maturation, testing, instrumentation, statistical regression, selection, experimental mortality, direction of causal conclusion, and exchange of information.

5. Threats to construct validity occur amongst others in the context of missing precise definitions of constructs, mono-operation bias, mono-method bias, hypothesis guessing, social desirability responding, experimenter expectancies, omitting relevant levels of constructs, simultaneous influence of several independent variables, interaction between testing and treatment, and restricted generalizability over constructs.

6. Threats to external validity occur amongst others in the context of an interaction between selection and treatment, setting and treatment, or history and treatment.

Questions

3.1. Is it possible that an intervening variable is not a confounding variable?

3.2. Give reasons for the primacy of internal over external validity.

3.3. What is meant by a null result?

3.4. What can be done in case of low statistical power?

3.5. What is meant by “sensitization”?

3.6. What can be objected against the study design in Section 3.2.5?

3.7. Explain by means of an example in which way instrumentation may cause a selection effect.

3.8. Considering Section 3.3.1, how could causal conclusions be derived for the construct “pain”?

3.9. Make some proposals how the generalizability of the conclusions from the example in Section 3.3.3 might be raised.

3.10. How could the influence of social desirability responding be avoided?

3.11. In which way might be tested, whether a Rosenthal effect is present or not?

3.12. What is the aim of single-blind, double-blind, and triple-blind studies?

3.13. What can be done in animal studies to avoid wrong conclusions due to experimenter expectancies?

3.14. Give an example for an interaction between testing and treatment.

4 Control of Extraneous Variables

Experimental design involves the careful planning of a study such that it becomes possible to draw causal conclusions from the outcome which cannot be made implausible by alternative explanations, the so-called threats to validity. For this aim, common sense, i.e. elementary logic combined with a little bit of reflection, is sufficient. Unfortunately, however, one frequently observes that studies are being performed which are virtually meaningless due to their obvious shortcomings, because conclusions from the outcome of such studies will not bear a closer examination. Such studies must be openly criticized for more than one reason, be it on conferences where their results are presented or in journals with an obviously insufficient reviewing system where they are published: First, in such pseudo-scientific studies men or animals are stressed, impaired or even killed without any justification. Second, wrong results are being transmitted by the unfounded conclusions of such studies which may cause a lasting damage. Third, in each of these studies resources are spent which, thus, cannot be given to justified studies.

The **extraneous variables** in the title of this chapter are those causal variables in which the researcher is not interested in, but which, however, influence the dependent variable, i.e., which are confounded with the independent variable to be studied. They impair a valid causal conclusion and are thus a threat to internal validity.

4.1 Randomization

The technique of **randomization,** which was proposed by Fisher (1925) and others (cf. Section 1.6 and 1.7), is, without doubt, one of the most important control techniques of experimental design. Randomization not only guarantees that the assumptions for the evaluation of the data with statistical procedures are fulfilled (statistical conclusion validity), but it also has the effect that causal conclusions become possible. Without randomization plausible alternative explanations (threats to internal validity) will always exist, which have to be made implausible in tedious procedures by using additional control measurements and control groups. It is obvious, though, that this is not really possible, as these potential explanations are infinitely many.

The term randomization just means that an additional random element is introduced into the experimental setting.

4.1.1 Randomization and External Validity

The term randomization has more than one meaning in experimental design. The first meaning is that of a random selection of a sample of subjects from a defined **population** in order to examine it in the study. The **sampling** is performed **without replacement**, i.e. a subject cannot be selected twice. By contrast, in random **sampling with replacement** a selected subject might be reselected, in principle, for the same sample. By using the notion **random sample** one implies that before the sampling the probability of being selected has been the same for all subjects in the population.

Whether such a random sampling can actually be conducted, depends on the size of the respective population. If the population to be investigated comprises only patients suffering from a rare disease, for instance, and which are, in addition, all known, a different integer can be assigned to each patient. The numbers can be written on cards and the cards shuffled and put into a box. Thereafter, the experimenter draws the cards blindfold from the box until the fixed sample size is achieved.

If the population, however, consists of the inhabitants of a whole country, e.g. of 80 million people, the above procedure cannot be conducted for practical reasons. In this case a different number is assigned to each subject, e.g. an integer between 1 and 100 million. By means of a computer program **pseudorandom numbers** can be generated which take values from 1 to 100 million and have a uniform distribution, i.e. each integer has the same probability to be selected. Now, successively, pseudorandom numbers are being selected until the prefixed sample size is achieved. Integers which have been assigned to nobody or which have already been generated are eliminated and replaced by new pseudorandom numbers.

An advantage of this kind of random sampling without replacement is that those causal relations whose existence was proved for the random sample might be generalized to the total population to a certain extent (external validity). Due to the randomness of the selection, this generalization might, however, not be justified. Nevertheless, the probability of wrongly assuming a causal relation on the basis of a statistical test, which was performed for the random sample, is not larger than the fixed significance level α of the test for which an arbitrarily small positive number may be chosen in advance.

In most cases this kind of randomization is not feasible. If one wants to prove that the memory of men is better than that of women a random sample of 100 men could be selected from the population of all men and similarly a random sample of 100 women could be selected from the population of all women. The dependent variable "memory" could be operationalized by randomly selecting 20 syllables from a list of 400 nonsense syllables. The 20 syllables are written in a random sequence on a sheet of paper and each of the 200 subjects is made learn these syllables for the duration of one minute. Five minutes later each subject is asked to write down all the syllables he or she can recall. At least with respect to the very special aspect of the construct "memory" which is tested here, it would be possible to draw a conclusion about whether in the total population men have a better memory than women.

In practice however, the planned study cannot be performed and, therefore, the respective question cannot be answered. We neither have the population of all people on earth at our disposal nor can we draw random samples from the subpopulations of either all men or all women. Even if we do not regard all people on earth as the total population but only the inhabitants of a large city as far as known to the administration, such a study would fail due to subjects refusing to participate in the study, subjects who cannot be located, subjects who cannot participate, because they are too old or too young or illiterate etc.

If the subjects are not being selected randomly, but are recruited by means of posters, insertions etc. offering a financial recompense, the found effects can by no means be generalized to the total population, even if the samples comprise 1000 subjects or more. One always has to take into account that the obtained samples might differ from the total population, with regard to unknown, but essential characteristics, i.e. each effect which is detected may have been caused by selection. It avails little if one tries to achieve the same composition for the nonrandom samples as for the total population, because only few characteristics of the total population are known, e.g.

the distribution of age and gender. The undeterminable number of characteristics by which subjects may differ from each other and which has to be known for generating really representative samples, is not known at all. Only a random sample from the total population allows a generalization, as only in this way **representativeness** with respect to all known and unknown characteristics can be assured. Conclusions which are based on selected and not on random samples, can only be generalized to populations for which these samples are representative in each respect. Since there is no way to describe these populations to a sufficient degree of precision this theoretical possibility of a generalization is without practical consequence.

From the above argumentation, we can conclude that, in principle, it is not possible to prove the existence of differences between men and women with respect to memory and other constructs, at least not as long as gender cannot be assigned randomly to subjects in a sample.

4.1.2 Randomization and Internal Validity

In a second type of randomization in experimental design, different factor levels or combinations of factor levels are randomly assigned to subjects or vice versa. We have, e.g., a sample of 60 patients suffering from Alzheimer's disease. This sample may consist of all patients in a certain hospital, i.e. it is not a random sample from a population of such patients. We consider an independent variable with two levels. One level corresponds to a "memory training", in which patients learn techniques which could help them to memorize things. The other level corresponds to a "control condition", where the patients learn techniques which could help them to perform certain daily activities. If now 30 of the 60 patients are selected randomly and get the "memory training", while the remaining 30 patients get the "control condition", a randomization has been performed.

In practice, the random selection can be performed by assigning one of the integers from 1 to 60 to each patient, with the number being different for each subject. The integers are written on cards, the cards are shuffled and put into a box. Now, the experimenter blindfold draws 30 cards from the box and the corresponding patients are assigned to the "memory training", the remaining patients to the control condition.

The kind of randomization just described, which at the same time is the most frequently used kind of randomization, is the only method known by which **structural equality** of different populations which are to be compared can be guaranteed at least in the statistical sense. This is true, as not only all characteristics, which are measurable or known, but also the potentially infinitely many characteristics, which are neither measurable nor known are matched. Here, structural equality means that the samples of subjects to be compared do not differ systematically and at most randomly in any of the characteristics.

Structural equality is one of the three conditions which are necessary for causal conclusions to be allowed and which are integrated under the generic term **comparability** (Biefang, Köpcke and Schreiber, 1979, p. 8). A second condition is **observational equality,** which means that a characteristic should be recorded in the same way for all subjects. If technical devices are being used as measuring instruments for a longer period, it is necessary to check their calibration regularly. When observer ratings are being used it is important that the observers' reference point is not shifted.

The third condition is **operational equality,** which means that the realization of a factor level, i.e. of a specific treatment should be performed in the same way for all subjects of the same sample. This condition is not fulfilled, e.g., if a condition corresponds to a specific surgical treatment and the operating surgeon improves his or her performance with each new operation.

Apart from the three conditions of comparability given above, **representative equality** has also been demanded, i.e. the transferability of the effects detected to a structurally equal population. Such a transferability or generalization is trivial if we consider a fictitious population which has the same structure. In real populations this assumption can only be made if the respective samples are random samples from the population to which one would like to generalize. However, as discussed above, such selections are difficult to realize.

4.1.3 Randomization in Factorial Designs

The third kind of randomization can be regarded as a special case of the preceding one. If each subject undergoes not only one but several treatments, the samples, as we have already pointed out, are no longer characterized by factor levels but by combinations of factor levels.

The random assignment of subjects to these combinations can be performed analogously to the procedure in Section 4.1.2. E.g., if we have four combinations of factor levels and 32 subjects, we uniquely assign the integers from 1 to 32 to the 32 subjects, write the 32 integers on 32 cards, shuffle those cards and put them into a box. Then the experimenter draws 8 cards from the box with closed eyes, and the corresponding 8 subjects are assigned to the first combination. The next 8 subjects, which are selected in this way, are assigned to the second combination and in the same way the 8 subjects for the third combination are selected. By this sampling without replacement the 8 subjects of the fourth combination are given by the 8 cards remaining in the box.

4.1.4 Multiple Treatments

Up to now no heed has been given to the following problem: in some situations not all levels of a combination of factor levels can be active at the same time. This problem will be illustrated by some examples.

In a learning experiment the factor "number of nonsense syllables" with the two levels "7 syllables" and "14 syllables" as well as the factor "color of the cardboard on which the syllables are printed" with the levels "red" and "blue" are being considered yielding four combinations of factor levels: "7 syllables/red", "7 syllables/blue", "14 syllables/red", and "14 syllables/blue" (cf. Figure 4.1).

	Color	
Number	red	blue
7	7/red	7/blue
14	14/red	14/blue

Figure 4.1: Combinations of levels for the factors "number of nonsense syllables" and "color of cardboard"

In this example two factors are active at the same time with the consequence that we may assume that the chronological order of the presentation of the factors does not have any influence on the dependent variable (e.g. "number of syllables which are reproduced"). However, it is easy to find counterarguments to this statement as one cannot rule out that at least a part of the subjects first perceives the background color and then the syllables, while the opposite is true for others. This might result in a selection effect by which non-existing causal relations may be simulated and existing causal relations may be concealed without a possibility to control this effect. The extraneous factor "order of perception", which is not studied in the experiment, would be confounded with the experimental factors being "number" and "color" in a situation like this.

That this danger might really exist, can easily be seen if one assumes the colors to be chosen that shrill and to attract that much attention, that the syllables which are printed in black are observed only at second sight. Presumably this would yield the order of perception "color-syllables" for all subjects. However, if the colors are chosen that pale that differences between them are difficult to realize, presumably all subjects will show the opposite order of perception. One can assume that for all interim solutions with respect to the choice of colors an unknown percentage of the sample will exhibit one order of perception, while another percentage will exhibit the opposite order. A selection effect can arise depending on which subjects exhibit which order of perception.

If, in a learning experiment, the factor "number of nonsense syllables" with the levels "7 syllables" and "14 syllables" as well as the second factor "time between learning and recall" with the levels "5 minutes" and "10 minutes" are being considered, we get the four level combinations: "7 syllables/5 minutes", "7 syllables/10 minutes", "14 syllables/5 minutes", and "14 syllables/10 minutes" (cf. Figure 4.2).

	Time	
Number	5	10
7	7/5	7/10
14	14/5	14/10

Figure 4.2: Level combinations of the factors "number of nonsense syllables" and "time between learning and recall (in minutes)"

In this example the second factor (time between learning and recall) can only become effective when the first factor (number of nonsense syllables) has already been effective. A reverse order of the factors is not possible. Therefore, there is no danger that a reversal of the order might yield another outcome or that causal relations between the effect of learning (dependent variable, e.g. "number of syllables which are reproduced") and the two independent variables (factors) can be influenced by the order of application of the factors.

In another learning experiment we again consider the factor "number of nonsense syllables" with the levels "7 nonsense syllables" and "14 nonsense syllables". But this time we consider as a second factor the factor "number of syllables with meaning" with the levels "7 syllables with meaning" and "14 syllables with meaning". This yields the four level combinations "7 nonsense syllables/7 syllables with meaning", "7 nonsense syllables/14 syllables with meaning", "14 nonsense syllables/7 syllables with meaning", and "14 nonsense syllables/14 syllables with meaning" (cf. Figure 4.3).

	With meaning	
Nonsense	7	14
7	7/7	7/14
14	14/7	14/14

Figure 4.3: Level combinations of the factors "number of nonsense syllables" and "number of syllables with meaning"

Here, in contrast to the preceding example, no order of presentation of the two factors is imperative. Theoretically, both factors might be presented simultaneously by presenting a list of nonsense syllables and of syllables with meaning in a random order such that a different order may result for each subject. A disadvantage of this procedure is that a different selective proceeding of the subjects, while learning the list, might influence the effect of the two factors in a way difficult to control. For example, it might be that subjects try to learn the nonsense syllables first if short lists are being presented and the syllables with meaning if they have to learn long lists.

An alternative may be to use a fixed order of syllables, e.g. by presenting always first the nonsense syllables and then the syllables with meaning. However, here it is not clear to which extent the acquisition of the second presented syllables is influenced by the pre-experience with the acquisition of the first presented syllables.

Another possible kind of presentation would consider only one factor with the four levels "7 nonsense syllables", "14 nonsense syllables", "7 syllables with meaning", and "14 syllables with meaning" instead of the two factors (cf. Figure 4.4). Twenty-four subjects could be randomly assigned to these four levels by analogy to the procedure above.

7 nonsense syllables	14 nonsense syllables	7 syllables with meaning	14 syllables with meaning

Figure 4.4: Levels of the factor "learning task"

This kind of proceeding is always advisable if one is only interested in finding out whether 7 nonsense syllables are easier to learn than 14 nonsense syllables, whether 7 nonsense syllables are more difficult to learn than 7 syllables with meaning, whether 7 syllables with meaning are easier to learn than 14 syllables with meaning, or, finally, whether 14 nonsense syllables are more difficult to learn than 14 syllables with meaning. The remaining two possible comparisons (7 nonsense syllables with 14 syllables with meaning and 7 syllables with meaning with 14 nonsense syllables) yield outcomes which are difficult to interpret. They contradict the principle of isolation formulated by Wundt (cf. Section 1.5) since two conditions (number and meaningfulness of syllables) are varied at the same time. Because of this it is not clear which factor is responsible for an observed difference between the two samples.

If one wants to know, in which way the timely order influences learning under the different conditions, we no longer have to consider a factor with four levels or two factors with four level combinations, but three factors with eight level combinations. The third factor ("timely order of two conditions") has the two levels "nonsense syllables in phase 1" and "nonsense syllables in phase 2". The syllables with meaning are given during the other phase. The eight level combinations "7 nonsense syllables/7 syllables with meaning", "7 nonsense syllables/14 syllables with meaning", "7 syllables with meaning/7 nonsense syllables", "14 syllables with meaning/7 nonsense syllables", "14 nonsense syllables/7 syllables with meaning", "14 nonsense syllables/14 syllables with meaning", "7 syllables with meaning/14 nonsense syllables", and "14 syllables with meaning/14 nonsense syllables" are given in Figure 4.5, where the condition which is presented first is always given before the diagonal stroke, and the condition which is presented second is given after the diagonal stroke. With respect to randomization this means that the set of the original 32 subjects is not randomly partitioned into four groups of eight subjects each, but into eight groups of four subjects each. This means that the timely order is randomized in addition.

		With meaning	
Nonsense		7	14
7	Phase 1	7 nonsense/7 with meaning	7 nonsense/14 with meaning
	Phase 2	7 with meaning/7 nonsense	14 with meaning/7 nonsense
14	Phase 1	14 nonsense/7 with meaning	14 nonsense/14 with meaning
	Phase 2	7 with meaning/14 nonsense	14 with meaning/14 nonsense

Figure 4.5: Level combinations of the factors "number of nonsense syllables", "number of syllables with meaning", and "timely order of both conditions"

The design which we have just discussed has a pronounced disadvantage: on the one hand, it allows to find out, e.g., whether the difference of the effects of 7 or 14 syllables with meaning is the same in phase 2 irrespective whether 7 or 14 nonsense syllables are active in phase 1. For accomplishing this, comparisons between the 1st and 2nd or between the 5th and 6th level combinations, respectively, must be compared, where always the same condition is effective in phase 1. On the other hand, it is not possible, e.g., to study the effect of the condition presented in phase 1 on the comparison of the levels "7 syllables with meaning" and "7 nonsense syllables" in phase 2, as there exist no two groups for which we have the same condition in phase

1, while in phase 2 the two levels under consideration ("7 syllables with meaning", "7 nonsense syllables") are presented. Because of this we cannot decide even if an effect is being observed, whether it is caused by a difference in phase 1, by a difference in phase 2 or by differences in both phases.

	Phase 2			
Phase 1	7 nonsense	14 nonsense	7 with meaning	14 with meaning
7 nonsense	7 nonsense/ 7 nonsense	7 nonsense/ 14 nonsense	7 nonsense/ 7 with meaning	7 nonsense/ 14 with meaning
14 nonsense	14 nonsense/ 7 nonsense	14 nonsense/ 14 nonsense	14 nonsense/ 7 with meaning	14 nonsense/ 14 with meaning
7 with meaning	7 with meaning/ 7 nonsense	7 with meaning/ 14 nonsense	7 with meaning/ 7 with meaning	7 with meaning/ 14 with meaning
14 with meaning	14 with meaning/ 7 nonsense	14 with meaning/ 14 nonsense	14 with meaning/ 7 with meaning	14 with meaning/ 14 with meaning

Figure 4.6: Level combinations of the factor "learning task" from Figure 4.4 and the factor "timely order"

In order to perform any comparisons in phase 2 with an arbitrary but constant condition in phase 1 one needs a design with 16 level combinations as displayed in Figure 4.6. In this case the set of the originally 32 subjects has to be partitioned randomly in such a way that 2 subjects are assigned to each level combination. The design in Figure 4.6 is constructed by using the design with one factor and four levels from Figure 4.4 for both phase 1 and phase 2. In addition to the eight level combinations of the design in Figure 4.5 the eight level combinations "7 nonsense syllables/7 nonsense syllables", "7 nonsense syllables/14 nonsense syllables", "14 nonsense syllables/7 nonsense syllables", "14 nonsense syllables/14 nonsense syllables", "7 syllables with meaning/7 syllables with meaning", "7 syllables with meaning/14 syllables with meaning", "14 syllables with meaning/7 syllables with meaning", and "14 syllables with meaning/14 syllables with meaning" must be considered.

Of course, the randomization procedure of the timely order of treatments described here, is not restricted to two time periods but can be extended to an arbitrary number of time periods.

We learn about a further kind of randomization in Chapter 9 when discussing single-case experiments. In this case subjects are no longer assigned to different factor levels or level combinations but a randomly selected design is assigned to a subject.

4.1.5 Randomization and Ethics

Especially in connection with clinical studies there exists a long and controversial discussion whether randomization is ethically justified or not (e.g., Hill, 1963; Gilbert, McPeek, and Mosteller, 1977; Burkhardt and Kienle, 1978; Brewin, 1982; Miké, 1989; Royall, 1991). Not only many patients but also many therapists cannot agree with the idea that the decision for a certain therapy is not based on a rational decision of a doctor after a professional diagnosis but rather on an additional random experiment that has nothing to do with the medical indication and does not take the state of health of the patient in question into account.

Many authors argue that randomized studies can be used without ethical problems provided that nothing is known before such a study, at least with respect to preceding well-planned other studies, that indicates that any of the treatments to be compared is superior or inferior to any other of the treatments. Before the study, neither the doctor nor the patients should have any information about which of the treatments compared should be preferred or refused.

Often a standard therapy, whose effectiveness has been known for some time, is compared with a new therapy, which is supposed to be superior to the standard therapy. As long as the superiority of the new therapy has not been proven undeniably it is certainly doubtful, from an ethical point of view, to apply the new therapy to patients instead of the standard therapy. On the other hand, it would also be doubtful with respect to ethical considerations to continue to use the standard therapy, if a new therapy might probably yield better results.

As discussed above, randomization is the only known technique by which nearly all possible kinds of alternative explanations can be made implausible. Therefore, only a randomized study allows causal conclusions when the best therapy of several therapies is to be determined. In all studies which are not randomized one cannot rule out that the therapy which proved best in the study is actually inferior to other therapies. One possible reason for a bad therapy coming off well could be, e.g., a selection effect, because without knowledge of the conductors of the study, a subgroup of patients with good chances to be treated with success is assigned to the apparent best therapy. But this very therapy might have no effect at all or even a harmful effect for patients with low chances. Or, it might be that a subgroup of patients with low chances to be treated with success is assigned to that therapy which, in reality, is the best one, though the effect of this therapy is not being investigated for patients with good chances.

This argumentation is not refuted by the opinion, that the distribution of the states of health as rated by the doctors is similar for the different treatments. Each rating of a doctor is based on a limited number of measurable characteristics and cannot consider all the relevant but, even today, unknown characteristics, which might be of importance when the state of health is rated. An optimal uniform assignment of patients to therapies with respect to the state of health can only be achieved by an appropriate randomization. Therefore, in the case of unproven inferiority or superiority of single therapies with respect to other therapies, randomization is not only allowed from an ethical point of view but ethics even demand the use of randomization.

In particular in connection with clinical studies the laws concerning legal protection of patients have been improved considerably in the last decades. One result is that each patient has to be informed not only that he or she participates in a clinical trial, but also to which treatments and measurements he or she might be assigned. Since a patient has the right to withdraw his or her consent to participate in the study any time, a successful randomization is considerably complicated and it is difficult to avoid selection effects.

In order to find a solution to this problem, Riecken et al. (1974, p. 57) consider the following four stages:

1. Compilation of a list of suited subjects.
2. Getting the consent to participate in all measurements which are identical for experimental and control group.
3. Random assignment of the subjects to the treatment conditions.
4. Getting the consent to participate in the treatment condition the subject was assigned to.

Stages 1, 2, and 4 may yield a reduction of the sample sizes. Though stage 3 is just the randomization stage, the authors argue that randomization can be used at an arbitrary stage of the schedule above and that the later randomization is used, the smaller is the danger of plausible alternative explanations.

In a discussion dealing with the timing of randomization Riecken et al. (1974, p. 175) propose to randomize only if a sample of suited subjects has been found, i.e. people who have consented to participate in the study no matter to which treatment they are assigned to, after having been completely informed about the design of the study. It is easy to see that there is a danger in proceeding this way, namely that subjects who were informed about the possible alternative treatments withdraw their consent contrary to their former promise, if they are assigned to a condition by the randomization procedure, they do not find attractive. In this case selection effects cannot be ruled out.

An, at first sight, perplexing proposal for the solution of this problem, which is, however, plausible at second sight was made by Zelen (1979), who considered the comparison of a treatment and control condition in clinical trials. According to this proposal a sample of patients is randomly split up into a group A and a group B. All patients of group A get a standard treatment. The patients of group B are asked whether they accept the application of the experimental therapy. If they accept, they get this therapy (group B1). If they refuse they get the standard treatment (group B2) just as group A. The crucial idea of Zelen (1979) is that all patients of group B (**intention-to-treat** group), i.e. also the patients with a standard treatment (group B2) are compared with patients of group A.

This proposal looks very odd, at least at first, because in group B subjects with both treatments are mixed. It is obvious, that it becomes more difficult to prove the existence of treatment effects if the number of patients in group B who refuse the experimental therapy increases. If all patients in this group refuse this therapy, the comparison becomes futile. If no patient in this group refuses this therapy we have the conventional randomized two-group design.

The advantage of this kind of procedure is that randomization is fully effective such that, e.g., selection effects can be ruled out as alternative explanation. Such effects must be taken into account if, e.g., group A is only compared to group B1 which got the experimental therapy. If this experimental design proposed by Zelen

(1979) is used, found effects admit a causal interpretation. However, it is possible that existing effects are not detected in case of a high rate of refusal. In this case, Zelen (1979) argues that a high rate of refusal indicates that a therapy which lacks attractivity to such a high degree should not be introduced into a clinical trial, at least not for the time given.

By the way, Zelen also made a proposal for a completely different aspect of clinical research: in most clinical trials one has to assume that the patients to be treated enter the study at different points of time and the recruitment phase may be long. An ethical problem arises if one notices that at a given point of time one therapy seems to have a better effect for considerably more patients than an alternative therapy. Here, one might ask whether one should stop, for ethical reasons, assigning patients to the seemingly less effective therapy, because this is required by a randomization procedure. On the other hand, a temporary superiority of one therapy may be due to a random effect. This can only be decided at the end of the study. To overcome this dilemma, Zelen (1969) proposes an **adaptive design (play-the-winner rule)** in which the assignment of a patient to a therapy depends on the outcome of the study hitherto observed. Because for Zelen's (1969) approach selection effects cannot be ruled out, Wei and Durham (1978) additionally introduced the randomization principle (**randomized play-the-winner rule**).

The proceeding for this kind of design can be described as follows: two therapies, A and B, are to be compared. Put u cards (with $u \geq 0$) on which an A is written and u cards on which a B is written into a box after shuffling thoroughly. If a large number is chosen for u, the design adapts only slowly to the numbers of successes or failures, respectively, of the two therapies which are reported back. However, if a small number is chosen for u, the design will, in the beginning, react very sensitively towards differences in the numbers of successes or failures for the two treatments, which are reported back.

If a patient enters the study and must be assigned to a therapy, a card is randomly drawn from the box and according to the letter on the card the patient is treated by therapy A or therapy B. When this decision has been taken the card is put back into the box and all cards in the box are shuffled again. If the box contains no card and a decision must be made, which can happen in the starting phase if $u = 0$ is chosen, a fair coin (head corresponds to therapy A, tail to therapy B) is used or a fair die (an even number corresponds to therapy A, an uneven number to therapy B). As soon as the outcome of the corresponding therapy is known for a patient who has already been treated, additional cards are put into the box and a new shuffling is performed for all cards in the box. Here, four situations are distinguished:

1. The patient received therapy A and this was a success. In this case v cards with A and w cards with B are put into the box additionally.

2. The patient received therapy A and this was a failure. In this case w cards with A and v cards with B are put into the box additionally.

3. The patient received therapy B and this was a success. In this case w cards with A and v cards with B are put into the box additionally.

4. The patient received therapy B and this was a failure. In this case v cards with A and w cards with B are put into the box additionally.

Here, it is assumed that $v \geq w \geq 0$ holds. Obviously, $(v + w)$ cards are always added if for a patient the result of the corresponding therapy becomes known. It is a particular advantage of this design that the assignment of therapies to patients does not require that one knows the effect of the therapies for patients already treated, though, if this information becomes available it can be used to improve the selection procedure.

If $v = w = 0$ is chosen, the composition of cards in the box does not change and the therapies A and B are assigned with the same probability .5, just as when a fair coin is being used. In this case, we have no adaptive design which adapts to the known results of the two therapies at a given point of time. We get the same result for the choice $v = w > 0$.

Now consider the case of $v > w$. If we get a positive feedback for one therapy in the majority of cases, while at the same time we get a negative feedback in the majority of cases for the alternative therapy, the probability that the next patient is assigned to that therapy which at the given point of time has shown more successes is increased. This means that the design takes the known outcomes into account. The larger the difference between v and w the more sensitive the design reacts to the hitherto existing feedback.

4.2 Elimination and Blocking Off

Because **extraneous variables** are causal variables which are not of interest though they have effects on the dependent variable, one will try to eliminate them, if their existence is known before a study, e.g., due to a preliminary experiment. We consider, e.g., an experiment where subjects have to learn nonsense syllables. During a preliminary experiment one has observed that some subjects are irritated by a leaking water-tap in the laboratory. An **elimination** of this source of irritation can be done by repairing the water-tap or by turning the water off completely.

Another interference is possible, if subjects can look through a window in the laboratory on a street while they are learning. At least a part of the subjects may be diverted by optical and acoustic stimuli observed on the street. Because these extraneous variables most probably cannot be eliminated by the researcher, one can try to block off the subjects against these distractors. Such a **blocking off** might be performed by covering the window-pane with non-transparent and sound-absorbing material by which uncontrolled acoustic and optic stimuli are excluded at the same time. A better, though possibly more expensive solution, might be to make the subjects go into a laboratory without windows which is isolated against noise from the outside (a so-called **camera silens**). If both solutions appear to be too costly, one might cover the window panes with black cardboard (to block off fluctuations of daylight apart from other optical stimuli) and try to reduce acoustic distractions by the use of headphones.

An elimination or blocking off of extraneous variables is only possible if the existence of these variables is known before the beginning of the real study. However, this foreknowledge is not sufficient for eliminating disturbances caused by different learning histories of the subjects before the study, by differences in motivation, or by different responses to the situation. Similarly, effects caused by differences in age, gender, body weight, intelligence, educational standard, social class, etc. cannot be eliminated or blocked off. In practice these control techniques can be used only for a few extraneous variables which must be known. If, however, such variables have a strong influence on the dependent variable, these are two very efficient techniques.

After the elimination or blocking off of known extraneous variables a randomization, as described in Section 4.1.2, must also be used to neutralize the influence of unknown extraneous variables. If only a randomization is used though an elimination or blocking off of extraneous variables would be possible and appropriate, the proof of the existence of causal relations is rendered more difficult because larger sample sizes are needed to reach the same efficiency.

4.3 Constancy and Covering

If it is not possible to eliminate or block off known extraneous variables, i.e. causal variables which are of no interest, one can try to keep these variables constant. Such a **constancy** may consist of keeping the optical impression of the room constant for all subjects in the laboratory, i.e. furniture, devices, and other things which are present, the location of these things in the room, the color of walls, floor, windows, and doors, the lighting etc. Similarly, the room temperature and air moisture has to be kept at a constant level. In addition, the experimenter should always be the same and the devices used for the experiment should not only have the same appearance, but should also be completely the same as for their functioning. While the study is being performed the type and order of measurements and the influence of the experimenter should be the same for all subjects. Also the instructions for subjects or shaping-procedures for animals, respectively, should be kept constant, i.e. the influence of the experimenter should be standardized and in each respect be neutral. By using the same experimenter for all subjects one avoids that, e.g., different responses of subjects are caused by the differing age and gender of the experimenters. With respect to constancy of all conditions there is one important exception: the realizations of the different levels of the independent variable, i.e. the experimental conditions, as conceived by the experimenter, should not be kept constant, but should differ in the planned way.

Human beings and animals which participate in the experiment should be homogeneous in all measurable variables, e.g. in age and gender, at least as far as possible. One thus avoids e.g., that an experimenter with given age and gender has a different influence on subjects differing in these variables.

It is not always possible to keep the extraneous variables constant. In many electronic devices which are used, e.g., for measuring dependent variables or for performing and controlling an experiment, ventilators are integrated which start running only after the device has been used for some time, if the temperature is higher than given temperature limits within the device. The starting and running of such a ventilator causes noise. Similar problems can occur with an air-conditioning plant which is controlled by a thermostat. This may start running as soon as the room has warmed up after some time, e.g. due to the subjects participating in the experiment. Other acoustic disturbances can occur due to a printer, the opening of a door, moves of the experimenter etc. Since even weak acoustic stimuli in an environment with just a few of these stimuli may influence the behavior of the participating subjects one can try to prevent the perception of these weak stimuli by **covering** them by a constant stronger stimulus. In the example above, a noisy ventilator which is always running could be used as a background noise to cover those noises which occur from time to time.

The term **covering** which we use here is not the conventional term used for this control technique which is often utilized in behavioral research. Usually the term **masking** is used instead (e.g. in Kling, Horowitz, and Delhagen, 1956). Considering the discussion in Mackintosh (1977, pp. 491-492), in particular with respect to the comparison of masking and overshadowing, we do not think that we should use the term masking in our context, because we assume that the stimulus which is covered should have no effect at all, and this is contrary to the definition of masking. Also, the term **overshadowing** which was introduced by Pavlov (1927, p. 141) has another meaning (cf. Mackintosh, 1977, pp. 492-497).

Of course, one cannot rule out and even must assume that the covering stimulus influences behavior. But as this influence is the same for all levels of the independent variable there is hope that the potential causal relations which are investigated are not influenced by covering. Such an assumption is not justified, if the covering stimulus has different effects for different levels of the independent variable. Assume, for example, that in a learning experiment nonsense syllables are presented visually via a screen on the one hand and audibly via loud-speakers on the other hand. If the covering auditive stimulus is too strong, the syllables which are presented audibly cannot be perceived. A control of extraneous variables by elimination, blocking off, and constancy should be preferred, therefore, to a control by covering as far as possible.

In addition to constancy or covering, respectively, of known extraneous variables a randomization, as described in Section 4.1.2, is necessary. Only in this way the influence of unknown extraneous variables can also be controlled. If only randomization is used and a possible and appropriate constancy or covering is dispensed with, the finding of existing causal relations is rendered more difficult.

4.4 Matching and Blocking

Randomization, elimination, blocking off, constancy, and covering are **global control** techniques. By contrast, matching or blocking, respectively, is a **local control** technique. Besides, **matching** or **pairing** is a special case of **blocking**. The difference is that the term matching is used, if an independent variable with only two levels is considered, while the term blocking is used, if the independent variable has more than two levels.

For the technique of constancy in Section 4.3, amongst others, the homogeneity of the subjects was required. Here, obviously, the two following problems arise: First, it might be quite difficult to find a sufficiently large homogeneous sample. Second, the generalizability of the results may be questioned the larger the homogeneity. A solution to both problems is local constancy. Here, one does not try to achieve homogeneity for the whole sample, but only for parts of it.

The proceeding is the following: First, observable variables are identified in preliminary studies. These are the more important the higher their influence on the dependent variable. These are the so-called **block** or **matching variables**. The amount of influence of such a variable on the dependent variable is sometimes measured by the absolute value of the correlation between the extraneous and the dependent variable. Here, "correlation" is a standardized numerical measure for the strength of a linear relationship between two variables.

Next, subsamples of subjects, which are called blocks, are formed. A **block** consists of subjects which are as similar to each other as possible with respect to the levels of all of these known extraneous variables. For a single independent variable a block must contain at least so many subjects that at least one subject can be assigned to each level of the variable. As a rule, the number of subjects for a block is chosen in such a way that it is a multiple of the number of levels. Then the same number of subjects can be assigned to each level. If the independent variable has only two levels, and if only one subject is assigned to each level, one gets pairs (so-called **statistical twins**) as a special case of blocks and uses the term **matching** instead of **blocking**. If homogeneous subjects from a block are assigned to the different levels of a factor in this way one says that the corresponding subjects have been parallelized. If two or more extraneous variables are to be controlled, the set of all possible level combinations of the different extraneous variables is considered instead of the set of levels of one extraneous variable.

The proceeding will be illustrated by an example. The independent variable is the variable "modality of presentation" with the two levels "presentation via headphone" and "presentation via screen" of 10 nonsense syllables for the duration of 10 seconds. The dependent variable is the "number of correctly reproduced syllables after 60 minutes". As an extraneous variable we consider "age" with the 16 levels "11-15 years", "16-20 years", "21-25 years", "26-30 years", "31-35 years", "36-40 years", "41-45 years", "46-50 years", "51-55 years", "56-60 years", "61-65 years", "66-70 years", "71-75 years", "76-80 years", "81-85 years", and "86-90 years". Two subjects are sought for each class of age, who agree to participate in the experiment. By this we get pairs of subjects which have been parallelized with respect to the corresponding class of age. Within a pair, the class of age is locally constant. In other words the subjects were matched with respect to the matching variable "age".

In order to control the potentially infinitely many extraneous variables which have not been kept constant, it is necessary to randomly assign one of the subjects within each pair of subjects to the screen condition such that the other inevitably is assigned to the headphone condition. This yields **randomized pairs**. The respective proceeding is a **local randomization** which corresponds to the proceeding in Section 4.1.2 for randomization with the exception that it is restricted to a randomization within subsamples.

This example shows several problems which are typical of matching and blocking. First, it is not clear which of the potentially infinitely many matching variables are to be considered. Thus one should know—though this will never be the case—which extraneous variable has the greatest effect on the dependent variable. If one has found in a preliminary study that an extraneous variable, e.g. age, has an exceptionally high positive or negative correlation (i.e., a strong linear relationship exists) with the dependent variable (e.g., the number of correctly reproduced syllables), this is an evidence for the usability of this extraneous variable as a matching or block variable. If we find now in the class of age "41-45 years" 20 subjects who agree to participate in the experiment, but only one subject in each of all the other classes, matching is difficult to perform even though we possibly have a high negative correlation between the matching variable (age) and the dependent variable (memory performance). Using a pragmatic approach, we might divide the 20 subjects of the age group "41-45 years" randomly into two subsamples of 10 subjects each. One of the subsamples is randomly assigned to one experimental condition. Then, the other subsample is assigned to the remaining condition. By this proceeding we get so-called **randomized blocks**. Underneath the age class "41-45 years" we still have six age classes, each with one subject. By pooling neighboring classes we can get three randomized pairs. Above the age class "41-45 years" we have nine age classes. By not considering the last age class "86-90 years" with one subject in the design, and by pooling neighbor classes, we get four additional randomized pairs.

If the independent variable had had six instead of two levels, e.g. six different memory training methods, each randomized block should have contained at least six subjects of an age group, because only then it had been possible to assign at least one subject to each experimental condition. It is sensible to have each block not only fulfill this minimum requirement, but also have each block contain an integer valued multiple of 6 subjects, e.g., 6, 12, or 18 subjects. Under these circumstances it is possible to randomly assign exactly 1, 2, or 3 subjects for each block to each experimental condition.

In such a situation one has to take into account that many available subjects cannot be considered in the experiment, because the blocks corresponding to the age classes cannot be filled up. If age classes are pooled as described above to evade this problem, the blocks may become so heterogeneous that blocking does not make sense. This problem is still aggravated if not only one but, simultaneously, more than one extraneous variable is considered for forming pairs or blocks. If we consider not only the extraneous variable "age" with 16 age classes in the example above, but in addition the extraneous variable "gender" with the two levels "male" and "female", we get a total of 32 level combinations. In order to perform blocking here, each of the 32 blocks must contain at least 6 subjects (in view of 6 training methods) which are homogeneous with respect to age and gender. If we take into account, e.g., that men have a lower life expectancy than women, it is possibly difficult to form a block with 6 men of the age class "86-90 years". But it might also be difficult to find sufficiently many subjects for other combinations, in particular, as we need for the 32 combinations and 6 levels of the independent variable a total of 192 subjects.

If not only the effect of one factor, i.e. of one independent variable, but the effect of several factors on the dependent variable is being investigated, level combinations of several factors must be considered instead of the levels of one factor. As the number of such combinations increases rapidly with the number of factors, matching or blocking becomes more and more difficult. In such cases, a pragmatic approach consists in confining oneself to such level combinations of the block variables for which enough subjects are available.

The difficulty to find matching or blocking partners, respectively, results in restricting sampling to certain subpopulations called **strata**. However, in this case the generalizability of results may be reduced because causal relations may be found which are valid only for certain strata. Furthermore, the possible loss of subjects from strata, for which only a few subjects are available for a blocking, may require a very large initial sample size.

A further problem consists in the selection of suited matching or block variables, since the correlation criterion described above cannot be used for qualitative data, e.g. occupation. Here, one has to test in a preliminary study whether the dependent variable changes more or less systematically with the levels of the extraneous variable.

If one does not use a suited extraneous variable for matching or blocking, this may have the effect that existent causal relations are not detected. However, if an unsuited matching or block variable is used, i.e. a variable which has no or only small influence on the dependent variable, an inefficient control procedure is being applied. This has disadvantageous effects on the interpretation of the outcome, if an evaluation procedure is used which takes the formation of pairs or, respectively, blocks into account. In this case, statistical power is lost in comparison with the usually employed statistical tests, i.e. effects that do exist in reality are detected with a smaller probability. This danger is not only present in case of completely unsuited matching or block variables which have no influence on the dependent variable considered, but it also occurs, though to a less degree, for matching or block variables with a low reliability, i.e. with a high error variance of measurement. If, e.g., anxiety is considered as an extraneous variable and those scores are used for matching or blocking which are exhibited by subjects in an anxiety questionnaire, one has to assume that due to a non-optimal reliability of the questionnaire subjects which differ very much with respect to their true anxiety are classified into the same pair or block. Therefore, the aim of local control is not achieved.

Sometimes the following artifice is used to overcome this problem: only **extreme groups** are used for matching or blocking. This means for our example that only subjects with high or low anxiety scores can participate in the study, while subjects with scores in the intermediate range of the scale are not admitted to the study. This procedure obviously restricts the generalizability of any conclusions, since an important part of the population is excluded. Further, if, e.g., a third of the originally tested subjects is not considered, it might be a more efficient proceeding to dispense with matching or blocking.

Another problem might occur, when matching is used, namely the so-called **overmatching** (Biefang, Köpcke, and Schreiber, 1979, p. 19). In this situation an independent variable does not influence a dependent variable directly but rather indirectly via a third variable. If this third variable, which is located between the independent and the dependent variable in the causal chain, is chosen as a matching variable, i.e. is kept constant within the strata defined by this variable, an influence of the independent variable on the dependent variable cannot be observed, though such an influence does exist in reality.

A simple example may illustrate this: 100 patients are assigned to each of two cancer clinics A and B. At the end of the treatment the perceived quality of life is measured by a questionnaire for the 200 patients. Researcher X finds by a direct comparison of the data from the two samples a significantly higher perceived quality of life for the patients from clinic A. Researcher Y performs a matching of the patients from clinics A and B before the evaluation. One matching variable is the variable "chemotherapy" with the levels "applied" and "not applied". Here, patients without a matching partner are not considered in the evaluation. Researcher Y finds no difference between clinics A and B with respect to the perceived quality of life.

This discrepancy between the two evaluation results can be easily explained, if we assume that chemotherapy is applied in clinic B on a considerably larger scale than in clinic A, and that the quality of life perceived by the patients is mainly influenced by the variable "chemotherapy" which was kept locally constant by researcher Y. Therefore, the conclusion of researcher X was correct, because the perceived quality of life for patients in clinic A was, actually, higher than for patients in clinic B, since chemotherapy was used to a lower degree in clinic A. A mistake in the proceeding of researcher Y was that randomization, i.e. the random assignment of patients to clinic A or B, was performed before and not after the matching. However, a randomization after matching might have encountered difficulties in view of the different strategies of using chemotherapy in both clinics.

One might have the idea that it is possible to perform blocking or matching and thus get randomized blocks, to pool those values of the dependent variable which correspond to the same level of the independent variable into the same sample, and finally to perform an evaluation procedure which pays no heed to pairing and blocking. In the learning experiment of the example above we would pool the measurements of those subjects who were assigned to the screen condition, irrespective of to which pair, defined by the same age class, the subjects belonged. Similarly, the measurements for the headphone condition would be pooled. Then, the two samples would be compared just as if no matching had taken place. In this case, the only purpose of matching would have been to get a better structural equality for the two samples with respect to age than with a global randomization.

A possible objection against this proceeding is that one cannot rule out an **interaction** between the independent variable and the matching or block, respectively, variable. Such an interaction might have the effect in the example above that better results are found for the younger subjects for the headphone condition, while the screen condition would yield better results for the older subjects. Depending on whether more younger or more older subjects must be left out of account for the matching procedure, as no appropriate matching partners are available, either the headphone or the screen condition might yield a better result. Thus, we would find an effect which depends on the more or less arbitrary choice of the matching variable, of the selection of the levels of this variable, and on the way the subjects are sampled.

We used an extraneous variable as a matching or block variable, i.e. a causal variable which influences the investigated dependent variable, but which is not investigated itself. This variable was to be controlled by local constancy of the extraneous variable, which is achieved by matching or blocking. The more the extraneous variable to be controlled influences the dependent variable the higher the positive or negative correlation of the two variables, if a linear relation is being assumed. An extraneous variable, which has exhibited a high positive or negative correlation with the dependent variable in preliminary studies is presumably an appropriate matching (or block) variable, if we leave out of account the appropriate matching (or block) partners which might be difficult to find.

McGuigan (1978, p. 206) made a proposal, which does not match the above considerations: the author reasons that the variable, which has the highest possible correlation with the dependent variable, namely the value of 1, is the dependent variable itself. Therefore, it might be a good idea to use the dependent variable itself as a matching or block variable. On first sight this idea seems to be rather odd, since a dependent variable is not a causal variable and, therefore, cannot be an extraneous variable which has to be controlled. A justification for such a proceeding might be that the dependent variable possibly measures a certain ability, e.g. "agility". If now the independent variable "training" has the two levels "method 1" and "method 2", we might perform a matching based on a **pretest** of the dependent variable such that subjects with the same extent of "agility" are compared with each other. This might facilitate the proof of a causal relation between the independent variable "training" and the dependent variable.

Note that in this example the dependent variable is, on no account, the extraneous variable but that the dependent variable is used as well for measuring the extraneous variable ("agility" before the independent variable "training" was effective) as for the measurement of the effect of the independent variable ("training"). Obviously, it is desirable for such a choice of the matching or block variable that the pretest does not have an influence on the posttest, which might occur, e.g., due to sensitization (cf. Section 3.2.3).

Considering the many problems which arise when using matching or blocking and which obstruct or even prevent the detection of causal relations, we agree with McGuigan (1978, p. 217) and Underwood and Shaughnessy (1975, pp. 61-62) that this control technique which was used very often in the past, should be treated with more scepticism nowadays.

4.5 Extraneous Variables as Independent Variables

One way to control extraneous variables is to regard them as **independent variables** in the experimental design. In most cases this will not concern variables which the experimenter can arbitrarily manipulate as this should always be the case for true independent variables, since these variables can be controlled by keeping them constant. However, there are also variables which, in general, cannot be manipulated, e.g. "age" or "gender".

In a learning experiment we consider the independent variable "modality of presentation" with the two levels "presentation via headphone" and "presentation via screen". In each trial we present 10 nonsense syllables for the duration of 10 seconds. We use the "number of correctly reproduced syllables after 60 minutes" as a dependent variable and the variable "gender" with the two levels "male" and "female" which in contrast to the independent variable "modality of presentation" cannot be manipulated as an extraneous variable.

We randomly select 20 subjects from a sample of 40 subjects without replacement and assign them to the headphone condition. The other 20 subjects are assigned to the screen condition. Apart from the dependent variable "number of correctly reproduced syllables" we additionally record the gender of the subjects. As gender cannot be manipulated it should actually be considered as a dependent variable. However, it would be difficult to consider gender as a dependent variable in this case, because we cannot assume that the modality of presentation has an influence on gender. If we consider "gender" as an extraneous variable which is to be interpreted as an independent variable, we get Figure 4.7.

	Modality of Presentation	
Gender	Headphone	Screen
Male	Male/Headphone	Male/Screen
Female	Female/Headphone	Female/Screen

Figure 4.7: Level combinations of the experimental factor "modality of presentation" and the extraneous variable "gender"

Note that the design in Figure 4.7 is different from a design based on matching or blocking in as far as that the random assignment to the levels of the factor "modality of presentation" was performed before gender was measured. Nevertheless, this is not really important. Similarly, it would have been possible to perform the random assignment separately for men and women in order to realize a blocking.

In contrast to matching or blocking, respectively, "gender" is regarded here as an independent variable whose effect on the dependent variable is to be investigated. When the outcome is being evaluated one does, therefore, not only ask whether the headphone and screen conditions differ in their effect on the number of correctly reproduced syllables (main effect of the experimental factor). One additionally asks whether there are differences in gender with respect to the dependent variable (main effect of the extraneous variable). A further important question is whether the difference between headphone and screen condition is different for both sexes or whether the difference between men and women has a different value for both modalities of presentation (interaction between the experimental factor and the extraneous variable).

The extension of a design by incorporating extraneous variables as independent variables seems to yield more information than matching or blocking. There, the parallelization had to be taken into account for the evaluation, but the matching or block variable should never be regarded as an independent variable in the evaluation.

The most important disadvantage in using extraneous variables, which cannot be manipulated as independent variables is that one must not give a causal interpretation of the main effect of the extraneous variable or of an interaction between the extraneous variable and the experimental factor. This could only be done in the not very realistic case of a true random sample from the population considered, as, otherwise, one cannot refute that the observed relations are solely due to selection effects. Apart from the not realistic case of a random sample one should refrain from treating extraneous variables which cannot be manipulated as independent variables and, in particular, for the case of matching or blocking these variables should not be included into any evaluation.

4.6 Replication

The use of replications is a technique, which is used to control systematic extraneous variables. A **replication** is the exact repetition of an experiment. If a causal relation has been detected by an experiment one cannot rule out that this relation has only been feigned by a random coincidence of various circumstances. Examples might be a nonrepresentative composition of a random sample or a random disturbance of the experiment.

If a causal relation, which is detected in an experiment, exists in reality, it is possible to predict the occurrence of the corresponding effect for any subsequent analogous experiments, too. If the same effect as in the first experiment is observed for an exact replication of the experiment, the confidence towards the existence of the effect is being considerably increased. This confidence nearly turns into certainty, if the same effect is found repeatedly in further replications. If, in one experiment, the probability that the effect occurred only by chance is given by .1, i.e. by 10%, the probability that this effect will occur by chance in each of three independent performances of the same experiment, is given by $.1 \times .1 \times .1 = .001$, i.e. by 1 per mill. Here, independence of the replications means that new random samples of subjects are drawn for each experiment.

Unfortunately, true replications are performed very rarely in behavioral research, though this would be a good strategy to assure that found effects are not merely due to chance. There seems to be a tendency to assume the existence of observed effects as given by a first and only study and to investigate by means of modified experiments other properties of these not necessarily existing effects, instead of confirming this existence by true replications. This wrong practice seems to be only marginally caused by the increased amount of time and costs necessary for replications. In this field of research, it is rather a fact, that replications of apparently known effects are not as highly esteemed by the scientific community as the proof of the existence of possibly not existing effects.

Only random but not systematically occurring disturbances can be controlled by replication. Therefore, replication can only be useful in randomized experiments. For example, in the investigations on talking and calculating horses, dogs, and cats in Section 3.3.7 many replications were used, and the described effects showed up repeatedly. In spite of this, the majority of scientists at that time, as well as probably each scientist nowadays, was convinced that the claimed effects could not exist, but were feigned by systematic disturbances. It is the great and admirable merit of Oskar Pfungst (1907/1977) to have detected the true causes of the observed effects by a large number of carefully planned experiments with many controls. Nevertheless, it would be extremely unfair and wrong to reproach Wilhelm von Osten or Professor Ziegler with fraudulent manipulations. In particular, it is known that Herr von Osten was honestly convinced that the effects he had found did really exist, and he unconditionally supported the investigations by Pfungst and others (cf. Pfungst, 1907/1977, p. 167).

This historical experience with the impossibility to control systematic disturbances by replication should not lead to the wrong conclusion that replication should not be used in general for the purpose of control. The replication of well-planned randomized experiments can, as discussed above, increase the evidence for found effects enormously.

The story of "Clever Hans", however, (which indeed must have been a very remarkable horse!) might have the effect that we develop a certain scepticism against the outcome of a whole direction in behavioral research. Here, we refer to the concept propagated by the American psychologist Burrhus Frederic Skinner (1904-1990) in his books and articles on how to study behavior, which was shared by many of his students.

At first sight, it seems that Skinner (1938, pp. 442-444) particularly rejected the use of statistical procedures in the analysis of behavioral data. However, the argumentation supporting his criticism is mainly based on the way in which he believed behavioral research should be performed. The control techniques he proposes and uses essentially consist in the application of replication, elimination, constancy, and repeated measures (with respect to the latter technique consider Section 4.12 and 6.3). In particular, randomization for the control of unknown extraneous variables is not considered by Skinner. A reason for this restriction to control techniques whose restricted efficiency has already been discussed by us to some extent is given by Skinner (1961, p. 141) as follows: according to his opinion, experiments should not be used to test theories, but to describe systematic changes of behavior. Here, non-observable causality is replaced by observable correlation (Skinner, 1938, p. 443).

In contrast to Skinner we had defined the task of an experiment (cf. Section 2.1) as to find out the causes for a certain behavior. Skinner is content with detecting observable relations without the intention to uncover hidden causal relations. In our opinion this kind of proceeding may lead to completely wrong conclusions with respect to laws determining a certain behavior. Skinner (1961, p. 141) believes that his kind of reasoning can yield results about the behavior of organisms which could change our society fundamentally, if they were introduced into education, commerce, industry, psychotherapy, religion, and government.

If, with respect to operant conditioning, only those investigations existed, which were performed in the way of Skinner's ideas and which only admit the description of certain correlative relations, it would be allowed to doubt the importance of these results. The observations of Wilhelm von Osten, Heinrich Ernst Ziegler, Karl Krall and others concerning the knocking language of horses and dogs were based on investigations which, just like Skinner's studies used insufficient control techniques and overinterpreted observed correlations without searching for the true causal variables. The effects reported by Skinner could also be interpreted as experimenter effects, because pigeons and rats like horses and dogs may transfer scarcely perceptible hints in the behavior of the experimenter into actions satisfying the experimenter's expectancies. That this is not only a purely academic discussion is demonstrated, e.g., in the studies of Rosenthal and Fode (1963) about the behavior of rats learning to discriminate two platforms and, in particular, in the studies of Rosenthal and Lawson (1964) about operant learning of rats in a Skinner box. The experimenter effects demonstrated in these studies show that the "Clever Ben", a calculating and writing rat, who masters the "lever language" in the Skinner box, does not have to be a mere fiction.

4.7 Balancing

A technique used for controlling supposed systematic extraneous variables is balancing. One tries to achieve structural equality by using this technique (cf. Section 4.1.2), i.e., to maintain that the conditions corresponding to the different levels of the independent variable (or, respectively, to the level combinations for several independent variables) do not differ in a systematic way.

For example, in a memory experiment the independent variable "modality of presentation" with the two levels "headphone" and "screen" is used. Two different laboratories are available, L1 and L2, two experimenters, E1 and E2, two experimental periods, A (morning) and P (afternoon), and 24 subjects of which 8 are men (M) and 16 women (W). If one now assumes that the headphone condition is always used with experimenter E1 in laboratory L1 in the morning and the screen condition is always used with experimenter E2 in laboratory L2 in the afternoon, difficulties may arise in interpreting the outcome of the experiment if differences between the two modalities occur with respect to the dependent variable.

It is difficult to disprove that such differences possibly are not due to the two conditions "headphone" and "screen" but are effects of the experimenter, the laboratory, and the experimental period. This would also be true, if such a constellation had occurred by chance as the result of an appropriately performed randomization with respect to experimenter, laboratory, and experimental period. Similarly, the interpretation of the outcome would raise doubt if 12 female subjects had been assigned to one of the experimenters and 8 male and 4 female subjects to the other experimenter. Here, again there would be objections against the interpretation of the outcome not only if this assignment had been performed systematically, but also if it had been the result of an appropriate randomization.

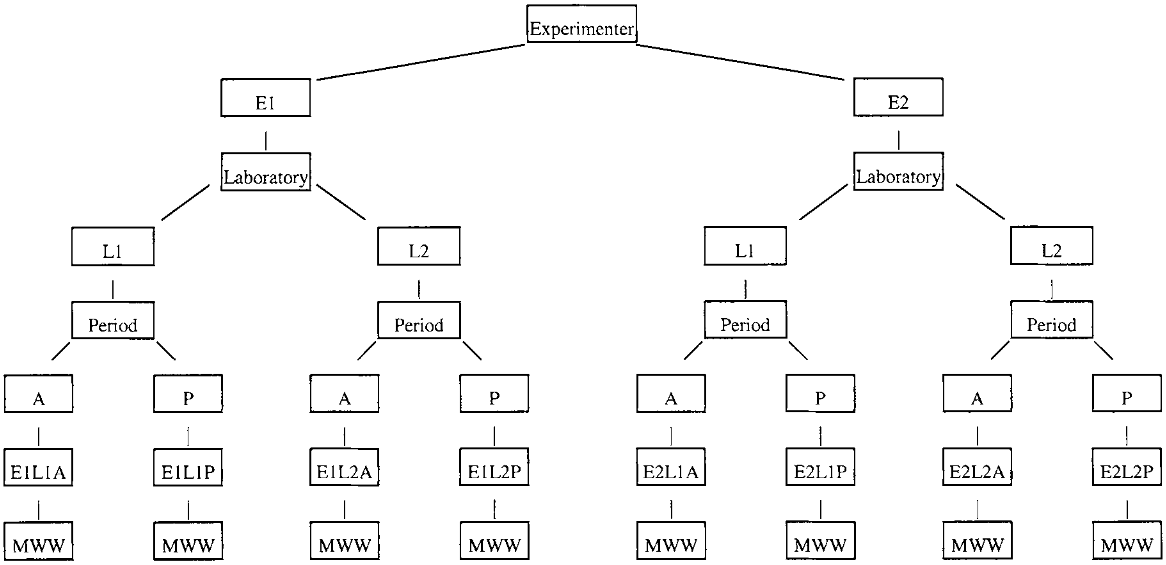

Figure 4.8: Possible experimental situations, i.e. possible combinations of experimenter, laboratory, and experimental period for the memory experiment, where one man and two women are randomly assigned to each situation.

For this example, **balancing** might be performed in the following way: The eight possible experimental situations, as depicted in the last but one line of Figure 4.8, are written side by side on a card. Then the integers from 1 to 24 are assigned to the 24 subjects such that all subjects get different numbers. The 24 integers are written on 24 corresponding cards. The 8 cards for the 8 men are shuffled and placed into a box. Similarly, the 16 cards for the 16 women are shuffled and placed into another box. Now, a card is randomly drawn from the first box without replacement and assigned to the first situation. The second card, which is drawn, is assigned to the second situation etc. up to the eighth situation. Then two cards are randomly drawn from the second box without replacement and assigned to the first situation. The next two cards are likewise assigned to the second situation etc. up to the eighth situation. The result is a design in which one man and two women are assigned to each of the eight experimental situations and where each of the possible combinations of experimenter, laboratory, and experimental period is realized. This design is called completely balanced with respect to experimenter, laboratory, and experimental period.

Since an odd number of subjects corresponds to each of the eight situations, it is not possible to distribute the two levels "headphone (H)" and "screen (S)" uniformly among the subjects within a situation. One way, to avoid interpretational problems here is to, first, shuffle two cards on one of which an H is written and an S on the other, for each of the eight situations. For one of the two women a card is drawn and hence the corresponding condition (H or S). The remaining woman gets the other condition. After that, eight cards (4 cards with an H and 4 cards with an S) are shuffled. Now, for each of the eight situations a card is randomly drawn without replacement and the corresponding condition is assigned to the man which had been assigned to this situation.

If it had not been 8 men and 16 women, but 20 men and 30 women instead, a **complete balancing** would not have been possible without further ado. In this case a box with 20 cards for the men and another box with 30 cards for the women would have been used. A first alternative might have been to randomly draw 4 cards from the first box and 6 cards from the second box without replacement and to exclude the corresponding subjects from the experiment. With the remaining 16 men and 24 women a complete balancing is possible, in which 2 men and 3 women are assigned to each of the eight situations.

A second alternative might be, to proceed first just like in the first alternative. However, at the end the 4 cards which were removed from the first box are replaced into this box which is empty now. Similarly, the 6 cards which were removed from the second box are now replaced into the empty second box. Then, the first box is restocked with 4 additional cards without a number and the second box is restocked with 2 additional cards without a number. Now, each box contains eight cards which are shuffled. The restocking is not applicable to a box which contains no cards after performing the first alternative. Now a kind of pseudo-complete balancing is possible where in the end to each of the eight situations 3 "men" and 4 "women" are assigned.

An advantage of this second alternative is that none of the 50 subjects has to be excluded from the experiment. A disadvantage is that we only have an **incomplete balancing** because, of course, no subject corresponds to cards without a number. This means that there are 4 situations with only 2 instead of 3 men and 2 situations with only 3 instead of 4 women. Nevertheless, the objections against the interpretation of the outcome of such an incomplete balancing certainly would be weaker than without any balancing.

The difference of balancing as opposed to matching or blocking, respectively, is that in the latter case the extraneous variables are rather related to the subjects, while for balancing the experimental situations are more important. That in reality the two control techniques cannot always be distinguished in such a clear way can be seen in the above example. The extraneous variable "gender" might be regarded as well as a block or as a balancing variable.

4.8 Counterbalancing

Counterbalancing is a control technique, which is usually used in order to control systematic disturbances, which might occur if the same subjects get more than one treatment. These disturbances occur because subjects do not respond to later treatments in the same way as to former treatments due to fatigue, exercise, decreasing motivation etc. Other causes are **carry-over** or **transfer effects**, i.e. after-effects of former treatments. It is not necessary that such after-effects are related to the immediately preceding treatment, but they may be caused, in principle, by all preceding treatments. Even more, one cannot assume that such after-effects are nothing but the sum of the effects of the preceding treatments. It is quite possible that it is not only important which treatments were applied before a given treatment, but also in which order these treatments were applied.

A special case of missing additivity of the effects is the occurrence of so-called **differential** or **asymmetric transfers**. This means that the transfer effect is not the same when the timely order of treatments is changed. **Irreversible effects**, i.e. effects, which cannot be removed, can serve as a trivial example for this. Asymmetric transfers are, above all, a threat to statistical conclusion validity, since the assumptions of additive models are violated.

Consider, e.g., that a neurotransmitter is studied which is assumed to improve the memory performance of rats. In an experiment the memory performance of a group of rats is measured before any treatment, then the rats get 1 mg of the neurotransmitter, the memory performance is measured a second time, then 2 mg are given and after that the memory performance is measured a third time. If a threshold for the effectiveness of the neurotransmitter exists at a dose of about 1.5 mg, no memory-facilitating effect will be observed if less than 1.5 mg are injected, while a lasting effect is observed, if more than 1.5 mg are injected. If the two doses are given in the described order, the second measurement shows no effect in comparison with the first measurement, while the third measurement shows an effect in comparison with the second measurement, because a total of 3 mg are effective in this case. If both doses are given in the reversed order, both treatments show an effect, because for the second measurement 2 mg are effective and for the third measurement 3 mg are effective. Different outcomes are observed depending on which order of treatments was chosen. Here, the asymmetric transfer is due to three different causes which work simultaneously: First, there is a threshold for the effectiveness, second, the effect is lasting and, thus, irreversible, and third, the effect is accumulative, i.e. different treatments with doses below the threshold may result in a crossing of the threshold and, thereby, produce an effect.

The term **complete counterbalancing** implies that to each possible order of the treatments the same number of subjects is assigned at random. For two treatments A and B the $1 \times 2 = 2$ orders A_1B_2 and B_1A_2 result. For three treatments A, B, and C $1 \times 2 \times 3 = 6$ orders $A_1B_2C_3$, $A_1C_2B_3$, $B_1A_2C_3$, $B_1C_2A_3$, $C_1A_2B_3$, and $C_1B_2A_3$ are considered. For four treatments there are $1 \times 2 \times 3 \times 4 = 24$, and for five treatments there are $1 \times 2 \times 3 \times 4 \times 5 = 120$ orders, etc.

As the number of possible orders increases rapidly with an increasing number of treatments, the number of required subjects also increases. Therefore, in **incomplete counterbalancing** one does not consider all possible orders any longer. Sometimes one only requires that

1. each treatment occurs equally often at each timely position,
2. each treatment is applied equally often before each other treatment.

A	B	C	D
B	D	A	C
D	C	B	A
C	A	D	B

Figure 4.9: Incomplete counterbalancing (with respect to the rows) for four treatments A, B, C, and D, where each treatment occurs exactly once at each timely position and exactly once before every other treatment

For two treatments A and B one cannot leave out any of the orders and the result with the two orders A_1B_2 and B_1A_2 is the same as for complete counterbalancing. For three treatments A, B, and C the two requirements above can only be fulfilled if all 6 possible orders are used, i.e. again no orders can be left out. For four treatments, e.g., the four orders $A_1B_2C_3D_4$, $B_1D_2A_3C_4$, $D_1C_2B_3A_4$, and $C_1A_2D_3B_4$ can be chosen from the 24 possible different orders and in this case both requirements are fulfilled. If these four orders are written one beneath the other the square in Figure 4.9 results. Another arrangement is depicted in Figure 4.10. However, the square in Figure 4.9 might be preferred to the rectangle in Figure 4.10, as only 4 orders are required. The square in Figure 4.9 has the property that each treatment occurs in each column exactly once, which follows directly from the first requirement. Further, each treatment occurs in each row exactly once. A square with these properties is called **Latin square**, because it can be depicted by using only Latin letters in contrast to more complicated designs. As shown in Figure 4.11, a Latin square does by no means have to fulfill the second requirement above, because the order A_1B_2 occurs twice (in order 1 and 4), but the order B_1C_2 only once (in order 1).

A	B	C	D
B	D	A	C
D	C	B	A
C	A	D	B
A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

Figure 4.10: Incomplete counterbalancing with four treatments A, B, C, and D where each treatment occurs exactly twice at each timely position and exactly twice before every other treatment

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

Figure 4.11: Latin square for the four treatments A, B, C, and D for which the second requirement is not fulfilled

For complete as well as for incomplete counterbalancing, and in particular for Latin squares, **randomization** can be performed by randomly assigning the treatments to the letters A, B, C etc. Furthermore, Latin squares for more than three treatments are no longer uniquely determined, i.e. cannot necessarily be transformed into each other by only replacing the original order of letters by another order or by interchanging rows or columns. In this case, a further randomization can be performed by randomly choosing one Latin square from the set of all essentially different Latin squares. In Figure 4.12 one of the possible alternatives to the design in Figure 4.11 is depicted.

A	B	C	D
C	A	D	B
B	D	A	C
D	C	B	A

Figure 4.12: Latin square for the four treatments A, B, C, and D which is essentially different from the Latin square in Figure 4.11, i.e. which results from this neither by interchanging rows or columns nor by interchanging the treatments A, B, C, and D

To avoid misunderstandings, note that Latin squares also exist for two (cf. Figure 4.13) and three (cf. Figure 4.14) treatments.

A	B
B	A

Figure 4.13: Latin square for the two treatments A and B

A	B	C
C	A	B
B	C	A

Figure 4.14: Latin square for the three treatments A, B, and C

If at each point of time two factors are simultaneously active (e.g., "kind of drug" and "dose of a drug") and if both factors have the same number of levels (e.g., the drugs A, B, C, and D with the doses α, β, γ, and δ) one kind of incomplete counterbalancing is a **Greco-Latin square** as depicted, e.g., in Figure 4.15. Latin and Greek letters are used in order to describe such designs. Here, each Greek and Latin letter must occur exactly once in each row and each column. Further, each combination of a Greek and a Latin letter must occur exactly once.

(α, A)	(β, B)	(γ, C)	(δ, D)
(β, D)	(α, C)	(δ, B)	(γ, A)
(γ, B)	(δ, A)	(α, D)	(β, C)
(δ, C)	(γ, D)	(β, A)	(α, B)

Figure 4.15: Greco-Latin square for two factors with four levels each

While complete and also incomplete counterbalancing might quite well control pure effects of exercise or fatigue, the outcomes may be difficult to interpret if treatments have different effects at different points of time, or if treatments do not only affect the directly following treatment, or if asymmetric transfers exist. E.g., Figure 4.11, where the order B_1C_2 but not the inverse order C_1B_2 occurs, shows that asymmetric transfers are not always controlled.

In spite of counterbalancing, the outcomes of designs with several treatments in succession never allow a causal interpretation. These difficulties arise already in the very common simple **crossover design** with two treatments A and B and two orders A_1B_2 and B_1A_2 (cf. Figure 4.13). For a more detailed discussion refer to Section 6.4.

In view of the considerable interpretational problems, which are typical of the technique of counterbalancing in designs with several successive treatments, this procedure should only be chosen if effects of order are actually being investigated. This technique must not be used in order to save subjects. If there is no interest in order effects, a simple design in which subjects get only one treatment should be preferred to counterbalancing, as a causal interpretation of the outcomes is facilitated. Furthermore, in view of the possible need of further control groups (cf. Section 6.4) it is not certain that a smaller number of subjects is required for counterbalancing in comparison with designs where only one treatment is applied. In any case, the stress experienced by the subjects is always higher for counterbalancing than for a design where only one treatment is applied to each subject.

Because the two terms balancing and counterbalancing look quite similar, there is a certain risk that these two control techniques are confused. In principle, **balancing** means that the levels of known or at least supposed extraneous variables are uniformly distributed over the different levels or level combinations, respectively, of the independent variables. **Counterbalancing,** by contrast, means that the levels of an independent variable are uniformly distributed over the levels or level combinations, respectively, of one or more known or supposed extraneous variables.

Assume, for example, that four therapies T1, T2, T3, and T4 are to be compared. The patients, who are available, are recruited from four clinics C1, C2, C3, and C4, where the severity of a disease is rated on a scale with four degrees, S1, S2, S3, and S4. In order to avoid that certain therapies accumulate in certain clinics and/or for certain degrees of severity, the Latin square in Figure 4.16 could be used, where the same number of corresponding patients is to be assigned to each of the 16 cells.

	S1	S2	S3	S4
C1	T1	T2	T3	T4
C2	T2	T4	T1	T3
C3	T4	T3	T2	T1
C4	T3	T1	T4	T2

Figure 4.16: Latin square for the four therapies T1, T2, T3, and T4 where each therapy occurs exactly once in each clinic and exactly once with each degree of severity of the disease

The example in Figure 4.16 shows that counterbalancing is a technique, which cannot only be used in designs with several treatments in succession to control for order effects, though we introduced this technique in exactly this context. If a therapy is not only determined by the used drug (D1, D2, D3, and D4), but also by its dose (d1, d2, d3, and d4), the Greco-Latin square of Figure 4.17 can be used.

	S1	S2	S3	S4
C1	D1 d1	D2 d2	D3 d3	D4 d4
C2	D2 d4	D1 d3	D4 d2	D3 d1
C3	D3 d2	D4 d1	D1 d4	D2 d3
C4	D4 d3	D3 d4	D2 d1	D1 d2

Figure 4.17: Greco-Latin square with the drugs D1, D2, D3, and D4 and the doses d1, d2, d3, and d4 for the clinics C1, C2, C3, and C4 and the degrees of severity of the disease S1, S2, S3, and S4

4.9 Blinding

The technique of blinding has already been mentioned in Section 3.3.6; disturbance effects, which occur because subjects form hypotheses about the true objects of the study in which they participate, can be controlled by this technique. In experimental laboratory studies, the subjects are usually not told which other experimental conditions exist and to which conditions they themselves are assigned to. This can have the consequence that subjects form their own hypotheses about the object of the study. Thus their response might be altered in a not controlled manner depending on which particular experimental condition was used. Subjects will often try to "help" the experimenter by producing outcomes of which they believe that they are desired. A further, possibly even stronger motivation is that the subjects try to appear as exceptionally competent, intelligent, healthy or "normal" in the sense of social desirability responding (cf. Section 3.3.5). If, however, a subject fears that recorded data might be used to curtail the claims of this subject with respect to insurance companies or any governmental institutions, it might happen that subjects intentionally try to present themselves in a negative way, e.g. as incompetent or ill. Both effects are possibly modulated by the age and the gender of the experimenter.

The best method to avoid such reactions of the subjects certainly would be to plan the experiment and the measurement of responses such that the subjects do not know that they participate in an experiment. Apart from ethical problems, this proceeding can usually not be used for practical reasons. It might, e.g., not be possible to randomly assign the subjects to the levels of the used independent variable under these circumstances or the dependent variable cannot be measured without the knowledge and consent of the subjects. For many studies, in particular in clinical research, law requires that the subjects are informed about the levels of the independent variable and the kind of measurements to be performed. In such cases, **blinding** seems to be an appropriate control technique, i.e. it ensures that subjects cannot detect to which level of the independent variable they are assigned. Obviously, in many experimental situations such a blinding is not possible because the different treatments can easily be discriminated. Even in drug research, where it is common practice to use blinding, a discrimination of the treatments by the subjects is often possible, for instance by observing the main and side-effects of the drugs.

In Section 3.3.6 the possibility that experimenters have certain expectancies about the outcome of a study was discussed as well as that information about these expectancies is transferred unintentionally and unconsciously to the participating subjects or animals. To avoid such transfers so-called **double-blind studies,** in which also the experimenter who is in direct contact with the subjects or animals does not know which treatment is applied to a given subject, are often used instead of the **single-blind studies** described above. The **code** for the actual assignment of subjects to the different treatments is controlled by the organizer of the study. In drug studies with patients, in particular, an ethical problem arises, as, if there are any complications, the doctor treating the patient can become active only after contacting the study organizer, because he or she does not know the treatment the patient got. Sometimes one tries to avoid such critical situations giving the doctor a means to decode the code in case of an emergency.

With regard to double-blind studies of long duration in particular, the problem might arise that patients are no longer willing to accept the uncertainty with respect to the kind of their treatment and will try to find the best possible treatment outside the study (Zifferblatt and Wilbur, 1978). There is also the possibility of "unblinding" or "blind breaking" if a patient or a treating doctor gets information about the true assignment of the treatment and thereby the double-blind condition will be violated. Finally, there is a danger that the experimenter tries to draw more or less correct conclusions about the applied treatment based on the reactions of the subjects. This again may induce experimenter expectancies, which can influence the outcomes.

Similar expectancies might also exist for the study organizer. Though one cannot expect any direct influence on the subjects, one cannot rule out that such an effect may influence the way of evaluating the resulting data, e.g., of how outliers or missing values are being dealt with. Here, one had better entrust the data to an evaluator who is not involved in the study and who gets the data in an anonymous form without communicating the treatments the different groups got, and without specifying the supposed directions of the effects. This is called a **triple-blind study**.

The discussion above shows that blinding may not be permitted for ethical reasons or may not be possible for practical reasons, because subjects or experimenters can directly discriminate the treatments or can draw conclusions with respect to the true treatment by considering the different effects of the treatments. No matter if these conclusions are right or wrong, they can lead to an increase of existing small treatment differences or can diminish large actual differences. In the same way non-existing treatment differences might be feigned to exist or the effects of existing treatment differences can no longer be perceived. In particular, if different treatments actually have different effects on the dependent variable, differing influences of the treatments on the attitudes of the subjects may as well help to detect actually existing differences or prevent such a detection. Such effects can also occur if different treatments have indeed the same effect on the dependent variable but different effects on other not considered dependent variables.

For example, the drugs A and B may have the same effect on the decrease of heart rate as dependent variable. However, if the drug A leads to a fur-like taste in the mouth, this possibly leads to a state of high anxiety of the patients by which the heart rate may be increased. In a double-blind study this might lead to the conclusion that drug B decreases the heart rate, while drug A increases the heart rate, if only the relative effect is considered.

The discussed interpretation problems, which arise when blinding is used, should by no means have the effect, that this technique is no longer used. The influences of the expectancies of subjects and experimenters are difficult to control if this procedure is not being used. It is our concern, to indicate that blinding, if it can be used at all, cannot always prevent misinterpretations.

4.10 Control Groups and Control Conditions

Researchers are usually interested in finding causal explanations of the following form: If treatment A is applied to subjects they exhibit behavior B. Conditional statements of this kind cannot be obtained by applying treatment A to a group of subjects and by then recording the behavior of this group. One cannot rule out that behavior B would also have been displayed if treatment A had not been applied. Behavior B, e.g. might be typical of subjects in the population, from which the group has been sampled. Or, behavior B was not caused by treatment A but by the specific experimental situation, e.g., by the appearance of the laboratory. Or, behavior B might solely be a reaction to a certain experimenter.

Furthermore, it might be that only a partial aspect of treatment A, which the experimenter does not want to investigate, e.g., a white cloak, has triggered the observed effect. Also emotional components, such as a conversation with the participating subjects or the handling (i.e. a gentle body contact) of the animals might be responsible for the occurrence of behavior B. If, however, a **control condition** is introduced which is totally identical to treatment A with the sole exception of the specific aspect of treatment A we wish to make responsible for behavior B, we can often unequivocally conclude from differences between the two conditions with respect to behavior B that the cause for the observed effect is the considered specific aspect of treatment A. Such a conclusion is justified in particular, if subjects are assigned to the two conditions at random, i.e. to the **treatment group** and to the **control group**. If the two conditions are applied to the same subject in a timely order and if differences are found between the conditions, it will be difficult to draw causal conclusions (cf. Sections 4.8 and 6.4).

4.10.1 Control Groups for Detecting Effects

As mentioned above, the observed effect of a treatment on a dependent variable can be assumed to exist if the subjects were randomly assigned to a treatment and control group. Here, it is admitted that the control group also has an effect, i.e. a difference between the two groups is ascribed to a difference of the effects of both conditions. This does, in general, not prove that a global treatment effect exists, but only that there is a specific effect of the treatment. This restriction of the interpretation of the outcomes is by no means undesired. That this kind of interpretation is justified is assured by guaranteeing that the only difference between the control and treatment condition is just related to this specific aspect of the treatment. The more differences between the two treatments are admitted, the more difficult it is to interpret differences of the outcomes.

4.10.2 Placebo Control Groups

The term **placebo** (Latin for *"I shall please"*) is a standard term, in particular in drug research. There it denotes a pseudo-drug, i.e. a substance for which no effect can be expected from a pharmacological point of view. This inefficacy is assumed if a placebo does not contain any agents for which pharmacological effects on the organism in question are known. Such placebos can be produced, e.g., in form of tablets, such that they do not differ in color, form, size, surface, solidity, taste, and odor, i.e., in all sensations from the real drug. The only difference is that the placebo does not contain that or those agents, which are to be investigated and which are contained in the drug.

If an experimental group gets a drug, while a control group gets a placebo under otherwise completely identical conditions, a better recovery rate in the experimental group shows that the additional agents contained in the drug are responsible for the recovery. However, if there is no essential difference in the recovery rates, this is an indication that any successes of the drug are not due to the contained agents but to other causes. Such causes are usually suspected in psychic reasons: E.g., the healing effect of a drug is ascribed to the experienced quasi-religious ritual where a medicine-man (the doctor) in ritual clothing (a white cloak) hands over a mysterious substance (the drug) to a help seeking believer (the patient). The patient must take this substance, after an emphatic and hope raising introduction (analogous to divine inspiration). For a recovery effect, which is caused by such a ceremony, it is obviously totally unimportant if the substance, which is being handed over, contains any special agents or not.

In double-blind studies in particular (cf. Section 4.9) the control group often receives a placebo. This is, however, not always appropriate, if one only wants to know whether a specific agent within a drug is in particular responsible for its positive effects. In this case it would be better to use a drug in the control group, which differs from the drug the experimental group receives only with respect to the presence of the agent of interest. In this case a drug would be applied in the control group too, e.g., a tested standard substance, for which a positive effect is known, i.e. which cannot be considered as a placebo.

It would not be correct to use the term placebo only in the context of drug research. E.g., the effect of psychotherapeutical measures in case of examination induced anxiety of students might be investigated. For this aim, 20 students of a sample of 40 students seeking help are randomly selected, which are randomly assigned to 20 specialized behavior therapists with at least five years of professional experience. The remaining 20 students are randomly assigned to 20 amateurs (e.g., probation officers, policemen, judges, parsons, barmaids). Each of the 40 relief workers has to dedicate one hour daily on five different days to his or her client. A specific effect of behavior therapy is assumed, if it turns out that considerably more students in this group pass a following examination than in the other group. Here, placebo is defined as a "therapy" which does not have a specific effect from a professional point of view.

For the time being, the above presentation of placebos has deliberately left out a number of problems. An important point is the ethical aspect in connection with placebos, which is discussed, e.g., in detail by Sissela Bok (1974). This article starts by describing a study of the side-effects of contraceptives where, in 1971, a control group of women received placebos in a clinic, with the consequence that ten of these women became pregnant.

A further problem deals with the question of what one does actually understand by a placebo. In most cases the definition of Shapiro and Morris (1978, p. 371) is cited. According to them a placebo is a therapy, which is used because of its nonspecific, psychological or physiological effect or which possibly has a specific effect which, however, is not effective in the considered situation. The authors founded their definition on the definition of "placebo" in the edition of Motherby's New Medical Dictionary from the year 1785 which they cite as follows:

> *"A commonplace method or medicine."*

The authors argue that this historical definition has been cited incorrectly for many years by replacing "or" by "of", which would mean a considerable restriction of the definition.

The definition of Shapiro and Morris (1978) is not undisputed. An alternative is the more complex definition by Grünbaum (1986), while Gøtzsche (1994) takes the view that there is no logically incontestable definition for the modern concept of placebo.

One of the first cases of a placebo control group in an experiment seems to be the study by Hollingworth (1912) about the effect of caffeine on quantity and quality of sleep. Hollingworth used sugar as a placebo, i.e. the control group got only sugar, the three experimental groups received doses of caffeine in addition to sugar. The experimental groups differed with respect to the times when they got the caffeine.

There are remarkable reports about placebo effects, i.e. about cases, where a placebo yielded essentially better results, at least for a subsample of patients, than drugs with a specific efficacy. Thus, Downing and Rickels (1980) report such a superiority of placebo over chlordiazepoxide for anxious outpatients in a double-blind study, while Harden et al. (1996) report the superiority of placebo, in this case saline, over ketorolac and meperidine in case of headache also in a double-blind study. That placebo effects can be observed not only for human beings, but also for animals, demonstrates a study by Dilsaver and Majchrzak (1990) where the injection of saline in rats yielded a higher core temperature in comparison with a control group without injection.

The question about what a placebo really is, becomes particularly important, if one observes that different substances which are all assumed to be without any effect prove to have different effects. Thus, Isaac and Isaac (1977) injected the placebo water into rats as well as the placebo saline. They not only found that both placebos resulted in a decrease of locomotor activity when compared to a control condition without injection, but they also found a difference between both placebo conditions.

While usually positive placebo effects are reported it is not astonishing that for men as well as for animals negative placebo effects have also been observed (Straus and von Ammon Cavanaugh, 1996, p. 317, p. 319). In such cases the term **nocebo** (Latin for *"I shall harm"*) is sometimes used.

Wolf and Pinsky (1954) report a study with 31 patients of the New York Hospital about the specific efficacy of mephenesin on subjective anxiety and tension and on their objective manifestations. They found that an improvement of the condition as well as toxic reactions occurred for both drug and placebo with similar percentages. Rosenzweig, Brohier und Zipfel (1993) pooled the results for placebo effects of healthy volunteers in 109 double-blind studies with a total of 1228 participants. They found adverse effects for 19% of the participants across all studies.

4.10.3 Control Groups for the Case that no Effects can be Detected

If the existence of a relation cannot be proved in a study, no statement is possible, in general. It is allowed, on no account, in such a case to declare that such a relation does not exist, because one cannot be sure that the sample size was large enough for proving the existence of an effect in spite of the variability of the data. Furthermore, one cannot rule out, in spite of an appropriately performed randomization, that, unfortunately, the random partition of the subjects had a result that did not admit the detection of existing effects though, of course, this will happen, in general, only with a small probability.

We will illustrate this by the following example: A food product has been developed which is pretended to yield a decrease in body weight. 20 subjects are randomly selected from a group of 40 subjects, which are nourished with the new product, while the other 20 subjects are nourished in the customary way. After seven days the subjects are weighted. If by the random assignment the 20 heaviest subjects were assigned to the experimental group (new food), possibly no effect is observed or even an effect opposite to the really existing effect.

Also, deficits in the operationalization of the levels of the independent variable may be the reason for existing effects not being detected. Using suited additional control groups such deficits can possibly be found out and considered in the interpretation.

E.g., one wants to compare a standard medicament with a newly developed drug. If no difference can be shown, it might be, that both drugs are effective but differ only to a small degree or not at all. On the other hand both drugs might not be effective and therefore no difference can be found. One can judge, which of these two cases is present, if a placebo control group is used in addition (cf. Section 4.10.2). If the outcomes for all these groups are similar, this is an evidence for the second case. However, if the placebo control group yields worse outcomes than the two other groups, this is an evidence for the first case. If one finds that the placebo control group yields better outcomes than the two drug groups or that the outcome of the placebo control group lies between the two other outcomes this admits conclusions with respect to the efficacy of the two drugs. These conclusions would not have been possible without the additional control group: In the first case we have evidence for an adverse effect of both drugs, in the second case we have evidence for an adverse effect of one drug and a positive effect of the other drug.

The use of an additional control group is of particular importance if one cannot be sure whether the experimental groups actually were assigned to different levels of the independent variable. If, e.g., the treatments consist of different kinds of training, one cannot rule out that none of these was effective, e.g., because the duration of each training was too short.

A completely different reason for not finding a difference between two treatments can be the use of a dependent variable by which it is not possible to measure the effects of the independent variable. By using an additional control group this kind of explanation can also be supported.

4.10.4 Yoked Control Groups

If a potential causal variable is influenced by the behavior of the subjects, it is difficult to give a causal interpretation of the outcome. To solve this problem, the use of yoked control groups has been proposed.

In the presumably first experiment with a yoked control group (Kling, Horowitz, and Delhagen, 1956) rats in a dark experimental chamber could flash up a weak light by pressing a lever, and each pressing of the lever triggered the light. In this experimental situation it is not clear whether rats press the lever more often because they are reinforced by the illumination of the dark chamber or if the on-off light condition causes a greater activity of the animals which is the reason that the lever is pressed with a higher frequency.

In the experiment, 24 rats were handled for 8 days to get used to the experimenter. The following 6 days, the rats were daily put into the chamber and the number of times they pressed the lever was recorded. In this phase, the light did not go on after a lever pressure. Based on the lever pressure rates, 12 pairs of rats were formed with nearly equal pressure rates and one rat of each pair was randomly assigned to the experimental group, the other one to the control group. During the following 11 days each lever pressure of a rat in the experimental group was answered by illuminating the chamber of this rat and, similarly, the chamber of its partner in the control group which was placed into another dark chamber, irrespective of whether this partner at this point of time also pressed on the lever or not. Lever pressures of rats in the control group did not induce an illumination of any chamber. If it was found out in this experimental design that the rats in the experimental group perform considerably more lever pressures than the rats in the control group, this indicates, according to the authors, that the going on of the light is the reason for a high rate of lever pressures and not a higher activity of the animals caused by the on-off light condition.

The term "yoked boxes" was introduced by Ferster and Skinner (1957, p. 36, pp. 399-407, p. 734). According to these authors "yoked boxes" are two chambers where in one chamber (yoked box) a reinforcement is given as soon as a reinforcement occurs in the other chamber. The intention is to control the frequency of reinforcement.

Other authors have extended this restricted understanding of **yoked control groups** by denoting as "yoked control" all arrangements which assign the consequences for the behavior of a subject also to a parallelized other subject. The first step to be performed should be a matching (cf. Section 4.4) by which a partner is sought for each subject which is parallelized with respect to one or more suited matching variables. Within each pair one partner is randomly assigned to the yoked control condition. As soon as the partner in the experimental group releases an external behavior consequence as a consequence of his or her behavior this consequence will also be imposed as an external intervention by the experimenter (or by a corresponding device, respectively) to the corresponding partner in the yoked control group irrespective of this partner's behavior. The idea is that differences between the two groups with respect to the dependent variable are due to the behavior of the partner in the experimental group and not to properties of the externally caused behavioral consequences.

Yoked control groups were used in many studies in psychology, in particular, in studies of the concept of "learned helplessness" which was introduced by M. E. P. Seligman (e.g., Seligman and Beagley, 1975). To avoid the wrong impression that yoked control groups were only used in former times in animal studies or for more theoretical problems and that this technique is nowadays no longer used, we give as a more recent reference an article by Freedman and Enright (1996). These authors use a yoked control group in an application-oriented clinical study for finding out whether "forgiveness" is a suited intervention goal for incest survivors.

Church (1964) demonstrated that the use of yoked control groups does not assure that causal conclusions can actually be drawn. According to Church differences between yoked control and treatment groups with respect to the dependent variable might be due to, e.g., individual differences of the members of a pair which, in spite of the appropriate randomization, may produce a pseudo-effect which occurs systematically.

Church's (1964) remark that the use of yoked control groups may induce wrong interpretations, might be illustrated for the study by Kling, Horowitz, and Delhagen (1956) considered above. We assume that in reality light is not a reinforcer for rats but that on-off light conditions increase the activity of rats and, hence, lead to an increase in the lever pressure rates. The amount of increase of activity induced by on-off light changes may be individually very different for the animals. If the experimental and the corresponding control animal are both strongly activated or if they are both barely activated, there will be no essential difference with respect to behavior (cf. Figure 4.18). If the experimental animal is hardly activated but the control animal rather strongly, the control animal experiences only a few on-off light changes. Therefore, this animal will also exhibit a low rate of lever pressures and, again, no essential difference in behavior results. If, however, the experimental animal is strongly activated but the control animal barely activated, the on-off light changes induced by the behavior of the experimental animal will have barely any influence on the behavior of the control animal, and the experimental animal will exhibit a higher lever pressure rate than the control animal.

In three situations no essential difference is found but in the fourth situation an enlarged activity of the experimental animal in comparison with the control animal is found. Due to this asymmetry we find a higher lever pressure rate in the experimental group across the groups, though light is not a reinforcer for the experimental animals.

	Activity Increase due to On-Off Light Changes		Activity of Both Animals
	Experimental Animal	Control Animal	
1. Subpopulation	High	High	Similar High
2. Subpopulation	Low	Low	Similar Low
3. Subpopulation	Low	High	Similar Low
4. Subpopulation	High	Low	High for the Experimental Animal, Low for the Control Animal

Figure 4.18: Outcomes of an experiment by Kling, Horowitz und Delhagen (1956) for yoked control pairs from four fictitious subpopulations according to the reasoning of Church (1964).

One might argue that one would have been able to causally interpret the outcomes if the authors had already used the on-off light changes in the baseline phase independent of behavior and if only animals with comparable activity changes had been paired. In the reported experiment the animals were paired with respect to their lever pressure rates during the last four days of the baseline phase and during this phase no light was given. However, such an alternative pairing is difficult to realize, because one does not know which matching variable, i.e. which measure of an activity change, might be used. Further, it is not clear, in which way an on-off light change in the preliminary phase, which is not tied in with lever pressures, may influence the lever pressure behavior in the main phase. Using such a design it is possible that one would get answers to other questions than to those originally posed.

Church (1964) demonstrates also for some other experiments which were described in literature that it is barely possible to give causal interpretations for outcomes of experiments with yoked control groups. Church proposes a design with independent groups as an alternative experimental design, i.e. with groups which are not yoked. They experience different intervals of delay between response and event in order to find out, whether the timely connection between an event (e.g., a light goes on) and a response (e.g., a lever pressure) is responsible for the observed effect (e.g., an increased lever pressure rate). If one investigates, however, whether the frequency of the events (e.g., the frequency with which a light goes on) is responsible for the observed effect (e.g., an increased lever pressure rate), it is possible, according to Church, to compare independent groups with different frequencies of events.

4.10.5 Expectancy Control Groups

In order to control experimenter expectancies (cf. Sections 3.3.6 and 3.3.7), Rosenthal (1966, pp. 380-400) introduced **expectancy control groups**. For example, no treatment is applied to such a group but one makes the experimenter believe that a treatment effect is expected with a specified direction.

	Independent Variable	
	Experimental Condition (E)	Control Condition (C)
Expectancy of An Effect by the Experimenter (O)	O – E	O – C
Expectancy of No Effect by the Experimenter (N)	N – E	N – C

Figure 4.19: Confounding of the independent variable with the expectancies of the experimenter (adapted from Rosenthal, 1966, p. 381)

Figure 4.19 reproduces Rosenthal's (1966) view of the coincidence of different experimental conditions with different experimenter expectancies. Here, according to Rosenthal, the combination O – C (experimenter expectancy of an effect though no treatment is applied) as well as the combination N – E (experimenter expectancy of no effect though a treatment is applied) correspond to expectancy control groups. If both corresponding control groups are used, Rosenthal calls this a **complete expectancy control**. If only one of the two groups is used the term **partial expectancy control** is used. By comparing the outcomes for independent groups of experimenters, who correspond to the combinations O – E and N – E or O – C and N – C, respectively, experimenter effects can be detected if they exist.

If one takes into account that a group of subjects corresponds to each experimenter and an equal number of experimenters to each of the four cells of Figure 4.19, it becomes obvious that such an investigation will be very expensive. However, Rosenthal (1966, p. 392) argues that an expectancy control group design can also be performed with only one experimenter who is used in all four cells of Figure 4.19. If only two experimenters are available, Rosenthal proposes to either expose both to all four combinations or to expose one to the combinations O – E and N – C, the other to the combinations O – C and N – E. In our opinion, the outcomes of expectancy control designs where one experimenter is exposed to more than one of the four combinations in Figure 4.19 are difficult to interpret, as one cannot rule out that transfer effects from one combination to the following ones exist. In an ideal design, and this is also Rosenthal's view, a large sample of experimenters is randomly assigned to the four combinations, where each experimenter is exposed to only one combination.

In Figure 4.20 an expectancy control group design with six combinations is displayed, which facilitates the interpretation of the outcomes as opposed to the Rosenthal design in Figure 4.19. The essential difference between the two designs is that one does not only make the experimenter believe that an effect is being expected but also that a directed effect is being expected. Thus, the comparisons of (O+) – E with (Oo) – E, (O+) – E with (O–) – E, (Oo) – E with (O–) – E, (O+) – C with (Oo) – C, (O+) – C with (O–) – C, and (Oo) – C with (O–) – C are suited for the detection of experimenter effects. However, there is a tradeoff between the better interpretability of the design and the costs: a design allowing an easy interpretation is rather expensive.

	Independent Variable	
	Experimental Condition (E)	Control Condition (C)
Suggestion of the Expectancy of a Positive Effect (O+)	(O+) – E	(O+) – C
Suggestion of the Expectancy of No Effect (Oo)	(Oo) – E	(Oo) – C
Suggestion of the Expectancy of a Negative Effect (O–)	(O–) – E	(O–) – C

Figure 4.20: Completed expectancy control group design

		Independent Variable	
		Experimental Condition (E)	Control Condition (C)
Experimenter	Subject		
Expectancy of An Effect (O_E)	Expectancy of An Effect (O_S)	$O_E - O_S - E$	$O_E - O_S - C$
	Expectancy of No Effect (O_{SN})	$O_E - O_{SN} - E$	$O_E - O_{SN} - C$
Expectancy of No Effect (O_{EN})	Expectancy of An Effect (O_S)	$O_{EN} - O_S - E$	$O_{EN} - O_S - C$
	Expectancy of No Effect (O_{SN})	$O_{EN} - O_{SN} - E$	$O_{EN} - O_{SN} - C$

Figure 4.21: Double confounding of the independent variable with the expectancies of experimenters and subjects (adapted from Rosenthal, 1966, p. 396)

As discussed in Section 3.3.4 not only the experimenter expectancies might influence the outcomes, but also subject expectancies. (This is true for human subjects. Whether animals also have expectancies is difficult to find out and can hardly be controlled!) According to Rosenthal (1966), this yields Figure 4.21, in which twice as many combinations as in Figure 4.19 must be considered. Effects of experimenter expectancies can be judged by comparisons of the combinations $O_E - O_S - E$ with $O_{EN} - O_S - E$, $O_E - O_{SN} - E$ with $O_{EN} - O_{SN} - E$, $O_E - O_S - C$ with $O_{EN} - O_S - C$, or $O_E - O_{SN} - C$ with $O_{EN} - O_{SN} - C$. Effects of subject expectancies can be judged by comparisons of the combinations $O_E - O_S - E$ with $O_E - O_{SN} - E$, $O_{EN} - O_S - E$ with $O_{EN} - O_{SN} - E$, $O_E - O_S - C$ with $O_E - O_{SN} - C$, or $O_{EN} - O_S - C$ with $O_{EN} - O_{SN} - C$.

By analogy with Figure 4.20, Figure 4.21 can be completed by assuming directed suggestions for experimenters and subjects, yielding altogether 18 possible combinations of conditions (e.g. (O_E+) – (O_S–) – E or (O_E–) – (O_So) – C). Due to the enormous number of experimenters needed, however, a corresponding expectancy control group design will not really work.

All the group comparisons which we proposed for the different expectancy control designs aimed at the detection of experimenter or subject expectancy effects. In general, this will also be the most interesting question for us. However, for each of these designs it is also possible to propose comparisons by which we can get information about a possible **interaction** between the independent variable and the experimenter or subject expectancy.

This will be explained for Figure 4.21. If, e.g., by comparing the groups $O_E - O_S - E$ and $O_{EN} - O_S - E$ an essentially larger difference than by comparing the groups $O_E - O_S - C$ and $O_{EN} - O_S - C$ is observed, this is an indication for an interaction between independent variable and experimenter expectancy. However, if, e.g., by comparing the groups $O_E - O_S - E$ and $O_E - O_{SN} - E$ an essentially larger difference is observed than by comparing the groups $O_E - O_S - C$ and $O_E - O_{SN} - C$, this indicates a possible interaction between independent variable and subject expectancy. But if it is observed, e.g., that there is an essentially larger difference between the groups O_E – O_S – E and O_E – O_{SN} – E than between the groups O_{EN} – O_S – E and O_{EN} – O_{SN} – E, this is evidence for an interaction between experimenter expectancy and subject expectancy. Finally, a second-order interaction between independent variable, experimenter expectancy, and subject expectancy is possible: there is evidence for such an interaction, if the difference between the differences for the groups O_E – O_S – E and O_{EN} – O_S – E or O_E – O_{SN} – E and O_{EN} – O_{SN} – E, respectively, is essentially different from the difference of the differences between the groups O_E – O_S – C and O_{EN} – O_S – C or O_E – O_{SN} – C and O_{EN} – O_{SN} – C, respectively.

4.10.6 Solomon Design

In a simple control group design one group of subjects is randomly assigned to a treatment condition, another group to a control condition. Then in both groups a dependent variable is recorded. If an essential difference with respect to the values of the dependent variable is found between the groups, it is concluded that the treatment has an effect on the dependent variable, i.e. that a causal relation between the independent and dependent variable exists.

In many studies this simple experimental design is completed by a control measurement of the dependent variable which is recorded before the treatment and the control condition are introduced. A possible argument for introducing such a **pretest** might be that one fears that treatment and control group might be systematically different with respect to the dependent variable even before the experimental conditions are introduced in spite of an appropriate randomization. If this should actually be the case, a causal interpretation of the outcomes may be difficult. If, e.g., the later treatment group has essentially larger pretest values than the later control group it may be difficult to ascribe larger outcomes of the treatment group in comparison with the control group, with respect to the **posttest** values, to an effect of the treatment. If, however, it is expected that the values of the dependent variable are larger in the treatment group than in the control group, but the posttest yields about the same values for both groups, this might be due to an unfavorable randomization, if the treatment group exhibits essentially lower pretest values than the control group.

The most frequently used argument for recording pretest measurements is, however, that a high variability of the subjects with respect to the dependent variable is assumed and it is hoped to eliminate at least a large part of this variability from the posttest scores by means of the pretest scores using statistical procedures. It is hoped that after such a statistical adjustment it is easier to prove the existence of effects. Without discussing here the arising statistical problems (cf. Section 4.13) it should be pointed out that after such a proceeding a causal interpretation of the outcomes is only possible, if several unfounded and also implausible statistical assumptions are made. Such assumptions are, e.g., that occurring effects are purely additive (e.g., if **difference scores** are used) or purely multiplicative (e.g., if **percentage change** scores are used).

If pretest scores are simple observations where subjects do not know that they are being observed, it can be assumed in many cases that the performance of the pretest has no influence on the posttest scores. Here, it is supposed that no information of the performance of the pretest attains the subjects.

As a rule, it must be assumed that the pretest is perceived by the subjects which can have the consequence that the posttest scores are influenced by the performance of the pretest. Here, there may be a direct influence of the pretest (cf. testing in Section 3.2.3 and instrumentation in Section 3.2.4) or an interaction between pretest and independent variable (cf. Section 3.3.10).

These problems will be illustrated by an example. It is to be tested if a new method of teaching children is more efficient than a conventional method. To this aim, 40 children are randomly split up into two groups, each with 20 children. The first group is taught by the new method, the second group by the conventional method. Before and after teaching, all 40 children have to complete a questionnaire by which their knowledge is tested.

An influence of the pretest on the posttest can be caused by a possible actualization of knowledge by reading the questions of the pretest or by a combination of knowledge actualized for different questions of the pretest. In both cases cognitive processes may be released by which at the time of the posttest even more knowledge is actualized. Such a release of cognitive processes by the pretest may be supported, possibly in different ways, by the different methods of teaching.

Even more obvious is the influence of pretests on posttests if, e.g., physiological measurements are considered which are performed for human beings or animals. Here, it seems obvious that subjects which are, due to the pretest, habituated to the measuring instrument may exhibit other scores in the posttest as it would be the case in the absence of a pretest.

An obvious conclusion from the considerations above is that reactive pretests should be avoided at all events in order to prevent that no causal conclusion can be drawn. If pretests are used, e.g., for achieving a certain habituation with respect to the measuring instrument, it should be tried to control the effects of the pretest. A natural way to do this are the control group designs which were proposed by Solomon (1949).

In Figure 4.22 the **Solomon four group design** with the three control groups N – E, P – C, and N – C is depicted. By comparing the posttest scores for the groups P – E and N – E or for the groups P – C and N – C, respectively, it can be detected whether the pretest has had an effect. If the size of the effect is different for both comparisons this indicates an **interaction** between the independent variable and the presence of a pretest.

	Independent Variable	
	Experimental Condition (E)	Control condition (C)
Pretest is Present (P)	P – E	P – C
Pretest is Not Present (N)	N – E	N – C

Figure 4.22: Four group design adapted from Solomon (1949)

If the control condition is defined by the absence of a treatment, it is not possible to decide for the group N – C whether a measurement is a pretest or a posttest score. In such cases Solomon (1949) dispenses with the control group N – C. From this the

Solomon three group design results with the groups P – E, P – C, and N – E. However, the outcomes of this design are considerably more difficult to interpret in comparison with the four group design: only the comparison of the groups P – E and N – E can be performed for detecting whether an effect of the pretest is present in the experimental condition. It can neither be detected whether such effects also exist under the control condition nor whether an interaction exists between the independent variable and the presence of a pretest.

4.10.7 Comparison Groups

The term **comparison groups** is used if subjects which are to serve as controls are not randomly assigned to the control condition but belong to the control group, because they have certain properties. If it is to be investigated, e.g., whether stationary hebephrenic schizophrenics show better results than normal subjects when perceiving the content of slides which are presented for a very short time we cannot form true control groups, because the properties "normal subject" and "stationary hebephrenic schizophrenic" had to be randomly assigned to subjects. If suited comparison groups are sought for a given group of stationary hebephrenic schizophrenics, one should make sure that the composition of the comparison groups is similar to that of the patient group with respect to gender, age, and educational standard. Here, suited comparison groups might be, first, a group of "normal subjects" from outside the clinic, second, a group from the clinic staff, third, a group of "normal patients" which are in-patients because of an organic disease or injuries due to an accident, and, fourth, a group of convicts. The second comparison group would be a control for ensuring that a difference between schizophrenics and normal subjects with respect to the used dependent variable "performance in perceiving the contents of slides" is not solely due to the special clinical environment. The third comparison group is being used in order to control whether any stationary treatment in a clinic can be responsible for the expected effect. The fourth comparison group might be used in order to control hospitalization effects. The first comparison group, of course, is the conventional control group. To judge whether an observed effect is a specific consequence of schizophrenia, it would be advantageous to have a fifth comparison group of manic-depressive in-patients, which are not schizophrenic at the same time.

In principle, it is not possible for comparison groups, i.e. in case of control groups, which are given and not the result of a randomization, to rule out selection effects (cf. Section 3.2.6). This makes causal conclusions impossible. One can only try to rule out a certain number of plausible alternative explanations, i.e. threats to validity, by many skillfully chosen comparison groups, though, naturally, it is never possible to provide enough comparison groups for the potentially infinitely many alternative explanations, one could find. Therefore, one should try to randomize, whenever this is possible, even if this means, e.g., that a very small sample of patients has to be divided into even smaller subsamples. Questions, which only allow comparison groups, had better not be investigated, as corresponding studies do never yield results, which allow unambiguous causal conclusions.

The terms **nonequivalent control groups** or **nonequivalent comparison groups** are also used instead of "comparison groups". Designs, which use control groups, which are not the result of randomization are also called **quasiexperimental designs**. This labeling of designs which do not admit a causal interpretation of the outcomes is misleading, because the term quasiexperiment is too similar to the term **experiment** which, nowadays, is only used for randomized studies.

4.10.8 Historical Controls

For animal studies in particular researchers sometimes argue that in the respective laboratory many experiments with randomized control groups have already been performed, for which the data are completely available. If, now, a new experiment is planned, in which the same dependent variable as in former experiments is measured, the question arises whether a new control group must be used or whether it is allowed to use the data of one or more **historical control** groups to lower—for ethical or economic reasons—the need for further animals. Similar questions arise, of course, also in studies with human beings.

In contrast to comparison groups (cf. Section 4.10.7), historical control groups are the result of a randomization though this took place in the past. Nevertheless, it is immediately evident that the use of such historical controls also makes causal conclusions impossible. Here, we cannot rule out threats to internal validity. This is not only due to selection effects (cf. Section 3.2.6), but also due to history (cf. Section 3.2.1) and instrumentation (cf. Section 3.2.4). Thus, one has to expect, that differences between experimental and control group with respect to the dependent variable cannot be solely attributed to the differences of the conditions introduced by the experimenter. Both groups can be samples from very different populations and the experimenters might have become far more experienced with respect to the performance of the experiment and of the measurements in the meantime. The experimenters of the original experiments, from which the historical control groups are taken, might also be completely or partly different from those in the new experiment. In experiments with human subjects one has to consider that the subjects in the historical control groups participated in the study in another historical and social context, than those in the new study.

As a consequence, historical control groups should never be used and one had better dispense with such a study. The argument that this technique might save animals does not really hold. As studies with such data, which were obtained under other conditions do not allow a conclusive interpretation of the outcomes of the study, the waste of animals due to the superfluous study must not be ignored.

4.11 Conservative Arrangement of the Levels of Extraneous Variables

A technique, which should make a causal interpretation of outcomes possible without controlling a known or supposed extraneous variable was described by Matheson, Bruce, and Beauchamp (1971, p. 24, pp. 78-79). The authors propose to arrange the levels of the considered extraneous variable such that the detection of a causal relation is rendered more difficult. If, in spite of this, such a relation were found, this would increase the evidence for the existence of the relation.

One of the two examples the authors use for illustrating their technique deals with an experimenter who would like to know whether cockroaches are able to learn. To study this question a sample of cockroaches is placed into a box, which consists of a dark and a light compartment. As soon as the cockroaches move into the dark compartment they receive a shock by the experimenter. After having received some shocks, the cockroaches should rest in the light compartment, if they are able to learn, and are expected to run into the other compartment, if the light compartment becomes dark and, at the same time, the dark compartment becomes light. As we have the preknowledge that cockroaches have a natural preference for darkness, one can conclude that the animals have learned and have not reacted in correspondence with a natural preference.

The experiment could be still improved by finding out for each cockroach in a baseline phase whether it spends more time in the dark or in the light compartment. In the experimental phase, animals with a preference for light would be shocked in the light compartment, animals with a preference for darkness would be shocked in the dark compartment.

In any case, the authors conclude that the animals are able to learn if they leave the compartment for which they had a natural preference before the study. It is argued that the detection of a causal relation has been made difficult by choosing a **conservative arrangement of the levels of the extraneous variable**. After all, the cockroaches must exhibit such a strong learning behavior that this has a larger effect than the opposite natural behavior.

One problem with this technique is that it can only take known or supposed extraneous variables into consideration. E.g., all cockroaches may have been placed into the dark compartment at the same time and all have been shocked at the same time. In the moment when the shock was applied, a heavy wagon passed outside the laboratory and the cockroaches perceived the vibrations of the building as a very aversive stimulus, but not the shock applied by the experimenter. The causal conclusion that the cockroaches had learned to avoid the shock by fleeing into the light compartment would be wrong.

This difficulty can be avoided by performing a true experiment instead of using the above technique. In this case the sample of cockroaches is randomly split up into an experimental and a control group. These groups are placed at the same time into two different boxes as they were described above. The experimenter electrically shocks the experimental group but not the control group. If the shock has no effect and the vibrations constitute an aversive stimulus for the cockroaches, we have the following outcome: because the vibrations of the building are perceived by both groups at the same time and with same strength, no behavioral difference of the two groups will be observed. Thus, the wrong causal conclusion would be avoided that the cockroaches learned to avoid the shock.

The argumentation used for the technique of the conservative arrangement of the levels of extraneous variables is similar to the argumentation of some experimenters after an investigation. If, e.g., pretests are available (cf. Section 4.10.6) the evidence for a predicted and found causal relation is often seen as especially convincing, if a large difference of the pretest scores between the control group and the experimental group has the opposite sign as the corresponding difference of the posttest scores. If, on the other hand, the two differences have the same sign, the evidence for the existence of a causal relation seems to be weaker. If one accepts this reasoning, one should first perform the pretest and decide afterwards which of the two groups will get the experimental condition and which the control condition. The assignment where the result for the pretest scores contradicts the direction of the prediction is chosen. However, this kind of reasoning does not seem to be a really good idea, if we think, e.g., of a possible effect of statistical regression (cf. Section 3.2.5).

In this context, we would also like to mention the argumentation, that the evidence for a found result is particularly high, if it contradicts the expectancies and interests of the experimenter who found the result. The oldest known example of this kind of reasoning was reported about 2500 years ago by the Greek historiographer Herodotos of Halicarnossos (484? - 420? B. C.) in the second volume of his opus "The Muses" with the subtitle "Euterpe". To avoid a citation of the Greek original, but to preserve the exotic and archaic atmosphere we cite here from the German translation by J. Chr. F. Bähr, which was published in 1866. There we find (Herodotus, 1866, pp. 21-23, spacing as given by Bähr):

> "Die Aegyptier waren, bevor Psammetichus König derselben geworden war, in dem Glauben, sie wären die ersten unter allen Menschen gewesen; wie nun Psammetichus zur Herrschaft gelangt war, wollte er gern wissen, welche Menschen wohl die ersten gewesen, und von dieser Zeit an glaubten die Aegyptier, die Phrygier seien vor ihnen da gewesen, sie selbst aber wären älter als alle Anderen. Als nämlich Psammetichus durch seine Nachforschung in keiner Weise zu ermitteln vermochte, welche Menschen die ersten gewesen, ersann er folgendes Mittel. Er gab zwei neugeborene Knaben gemeiner Leute einem Hirten, der in seiner Heerde dieselben aufziehen solle in der Art, daß er ihm gebot, keine menschliche Stimme vor denselben hören zu lassen, sondern abgesondert in einem einsamen Gemach solle er sie für sich liegen lassen and zur bestimmten Zeit Ziegen zu ihnen führen: hätten die Knaben dann mit der Milch der Ziegen sich gesättigt, so möge er weiter seine Geschäfte besorgen. Also tat Psammetichus and also ordnete er an, weil er wissen wollte, welchen Laut die Knaben, wenn sie über die Zeit des undeutlichen Lallens hinausgekommen, zuerst von sich geben würden. Und dies geschah auch. Denn nach Verlauf von zwei Jahren, während welcher der Hirt also that, wie ihm befohlen war, liefen, als er einst die Thüre öffnete and eintrat, die Knaben zu ihm and schrieen, die Hände ausstreckend: Bekos. Wie dieß der Hirte vernommen, verhielt er sich anfangs ruhig, als er aber, so oft er zur Pflege der Knaben kam, immer wieder dieses Wort hörete, da machte er sofort seinem Herrn die Anzeige and führte, auf dessen Befehl, ihm die Knaben vor; and als nun Psammetichus es selbst gehört hatte, suchte er zu erforschen, was das für Menschen wären, welche das Wort Bekos im Munde führten. Bei dieser Erkundigung erfuhr er dann, daß die Phrygier damit das Brot bezeichneten. Aus diesem Vorfall erkannten die Aegyptier, daß die Phrygier älter als sie wären, und gaben es zu."

In our translation this reads:

> "Before Psammetichos became their king, the Egyptians believed that they were the first of all people. When Psammetichos came to power, he wanted to know which people had been the first and from this time the Egyptians believed that the Phrygians had existed before them, but that they themselves were elder than all the other people. It was then that Psammetichos had the following idea, when in spite of all his investigations he was not able to find out which people was the first. He gave two new-born boys of the common people to a goatherd who had to bring them up in his flock in such a way that they would not hear any human voice but lay separately in a lonely room. At certain times he had to bring goats to them that the boys could drink their milk. Thus, acted Psammetichos and ordered that he wanted to know the first word, which the boys would utter when they had left the period of inarticulated babbling. And this was done. After two years during which the goatherd followed the order, one day the boys came to the door, when he opened it to enter the room, and cried holding out their hands: bekos.

> When the goatherd heard this, at first he did nothing. However, when again and again he heard this word, whenever he came to foster the boys, he informed his lord and executing his order he brought the boys before him. Now, when Psammetichos himself heard the word he tried to investigate which people used the word bekos. He learned that the Phrygians used bekos for bread. From this incident the Egyptians learned that the Phrygians are elder than they themselves and admitted this."

Here, an important argument, though it is not given explicitly in the text, is that King Psammetichos II of Egypt (reign: 595 - 589 B. C.) held the opinion that the eldest people were the Egyptians. By the study performed by him this expectancy was defeated what is then taken as strong evidence for the conclusion that the Phrygians are the eldest people.

By the way, subsequent to the above Herodotos reports the following (Herodotus, 1866, p. 23):

> "Also vernahm ich den Hergang von den Priestern des Hephästos zu Memphis; die Hellenen dagegen erzählen darüber mancherlei einfältige Geschichten, so auch, daß Psammetichus Weibern die Zunge habe ausschneiden and dann bei ihnen die Kinder aufziehen lassen."

In our translation this reads:

> "Thus, I heard the course of events by the priests of Hephaestos in Memphis. However, the Hellenes tell about this some foolish stories, e.g., that Psammetichos had cut off the tongues of some women which had to bring the children up."

This version, which was rejected by Herodotos, would have brought even more evidence for the conclusion drawn in view of a remark of the translator which is given in a footnote concerning the word "bekos" (Herodotus, 1866, p. 22):

> "Man wird hier unwillkürlich an den von den Ziegen ausgehenden Laut (B e k) erinnert, welchen die Knaben nachgeahmt."

In our translation this reads:

> "Here, one is involuntarily reminded of the sound (b e k) made by goats and which was imitated by the boys."

4.12 Repeated measures

If more than one measurement for one or more dependent variables is performed at the same subject the term **repeated measures** is used. On the one hand, this concerns situations, where only one level of the independent variable is considered for each subject. In the simplest case only one measurement before and after a treatment is performed. However, sometimes also several measurements are recorded before the treatment in a so-called **baseline** and in some cases several measurements are recorded after the treatment. In both situations we have a **time-series design** with **intervention**.

In other cases several levels of the independent variable are considered for each subject. Even if only one posttest score is recorded for each treatment, several measurements are obtained for each subject, thus yielding again repeated measures. This is also true, if in such cases in addition to the posttest a pretest is performed for each treatment, which has to be distinguished from a posttest of a possibly preceding treatment. Again it is possible to record more than one measurement before the first treatment (baseline) and similarly between the different treatments and after the last treatment. Then, a time-series design with more than one intervention results.

Researchers, who use such **within-groups designs**, i.e. employ repeated measures, declare that these designs have a lot of advantages compared to those between-groups designs where only one level of the independent variable and one posttest is considered for each subject. (It is obvious that there are also mixed versions of these two types of designs where several independent groups are used and where more than one measurement is recorded for each subject.)

In most cases, the following advantages of within-groups designs are quoted:

1. One needs a smaller sample size than for between-groups designs.
2. Each subject serves as its own control.
3. The expenditure with respect to time and effort when performing measurements and treatments is lower than for between-groups designs.

The first advantage is, of course, only realized if more than one treatment condition is considered for each subject. The second advantage results, because the error variance can be reduced to a larger extent than for matching and blocking because the subjects are not only similar with respect to selected matching or block variables, respectively, but they are identical with respect to all possible variables of this kind. The third advantage results because in particular for largescale technical measurements the time spent may be reduced considerably, if it is not necessary to adjust the measuring instrument before each new measurement.

As we learned in some of the preceding sections, the declared advantages do not exist in reality because the internal as well as the external validity of the possible interpretations is threatened. This concerns, with respect to internal validity, the threats by history (cf. Section 3.2.1), maturation (cf. Section 3.2.2), testing (cf. Section 3.2.3), instrumentation (cf. Section 3.2.4), and statistical regression (cf. Section 3.2.5). Because subjects must be available for a longer time for repeated measures as opposed to single measures, in particular, if between treatments (e.g., in case of drugs) **wash-out periods** are provided, which are introduced for the fading away of the immediately detectable effects, one has to reckon with experimental mortality to a higher degree (cf. Section 3.2.7), i.e. incomplete data and a drastic reduction of the sample size result very often.

Especially in studies with human subjects selection effects (cf. Section 3.2.6) must be expected because only a subpopulation of subjects with unknown characteristics will agree to participate in a study with a long duration. This threatens the external validity of the conclusions.

In Section 4.1.4 and 4.8 we have already discussed why designs with more than one treatment yield outcomes which are difficult to interpret. However, if one wants to investigate the mutual influence of several treatments this requires very expensive experimental designs, which will be discussed in Section 6.4, such that no subjects are saved. This has already been illustrated in Section 4.1.4 where two different factors were arranged in a timely order. Other problems when using designs with several treatments for the same subject have already been discussed: it is not always possible to present treatments in an arbitrary order because treatment effects need not to be reversible. Further, one cannot be sure that the observed effects are the same irrespective of the order of treatments. Even with only one treatment applied to a subject, problems in interpreting repeated measures may arise, e.g., if a pretest has been recorded (cf. Section 4.10.6).

Up to now an important argument against using designs with repeated measures has not been discussed: several measurements at the same subject must be considered as being dependent in the statistical sense. As a consequence the assumptions of the conventional statistical tests (cf. Section 3.1.2) are not fulfilled, as these require, in particular, independent measurements. The so-called "repeated measures analysis of variance" or time-series analyses are based on assumptions of which one cannot be sure that they are valid for a given set of data. Calculations and simulations have shown that even small deviations from the required assumptions can considerably increase the probability that apparent effects are "detected" which do not exist in reality.

All these arguments against the use of repeated measures as a control technique are well-known and have been published repeatedly. Nevertheless, even today these designs are frequently used, because the researchers lack the resources or the motivation for an appropriately planned study. Frankly speaking, the resulting data from such badly planned studies can often not be interpreted in a conclusive way, which means that resources have been wasted and an unethical attitude has been exhibited as far as patients or animals are concerned.

Sometimes people argue that in the nineteenth century psychophysics as well as memory research was based on long series of studies with very few subjects. In many cases this was single-subject research and, quite often, the experimenter and the subject were identical. Similarly, in the twentieth century many results in behavioral research were found by studying the behavior of three or four pigeons or rats over a period of months. For the above reasons the results of such studies should not be used as a basis for further research. These results should rather be checked using designs, which admit causal conclusions. Checks by means of true experimental designs of apparent results found with repeated-measures designs show, that deviating outcomes may result. Unfortunately, these necessary checks were only rarely performed.

E.g., in 1885 the German psychologist Hermann Ebbinghaus (1850-1909) published his opus "Über das Gedächtnis (About Memory)" which is a book considered to be fundamental for memory research. In this book Ebbinghaus described experiments which tended to last several months and which were performed with one subject only, namely Ebbinghaus himself (Ebbinghaus, 1885, p. VI, p. 35). Underwood (1957) reports experiments with outcomes which show that at least the results with respect to the relation between the amount of forgotten material and the length of time since learning, which was derived by Ebbinghaus (1885, pp. 93-109), are not correct. The found relation can mainly be explained by an interference with material which was learned before in the laboratory, i.e. the relation derived by Ebbinghaus is an artifact caused by the used within-subjects design.

Of course, one cannot rule out that within-subjects designs may produce the same results as between-subjects designs if certain assumptions are valid, which are difficult to check. However, there exist quite a few studies, which demonstrate that the use of both kinds of designs may produce different outcomes. We refer here only to the articles by Grice and Hunter (1964), Pavlik and Carlton (1965), Mellers, Davies, and Birnbaum (1984), and File (1992). If we keep in mind that, in general, only appropriately performed between-subjects designs admit a causal interpretation, it is obvious, which results can be trusted when using both kinds of designs for the same problem. This also explains the scepticism of many authors with respect to the use of within-subjects designs (Grice, 1966; Poulton, 1973, 1974; Greenwald, 1976; Knapp, 1982).

As already discussed in Section 4.1.4, problems exist which require repeated measures, because, e.g., the combination of effects of different consecutively applied treatments is investigated. The appropriate experimental designs, which admit a causal interpretation of the outcomes, are, as a rule, more expensive than the convenient within-subjects designs.

4.13 Statistical Adjustment

Researchers sometimes prefer to use no matching or blocking (cf. Section 4.4), respectively, or any other of the control techniques which were discussed, if they deal with known or supposed quantitative extraneous variables, but try to eliminate the influence of the considered extraneous variable by using a **statistical adjustment**. Because one always has to assume that data are altered by measurement errors—otherwise statistical procedures would be superfluous—no statistical adjustment can be perfect. One usually has to assume that an **underadjustment** results where the influence of the extraneous variable is not totally eliminated or an **overadjustment** where the influence of the extraneous variable is overcompensated. The general proceeding is to postulate a specified influence of the extraneous variable on the measurements in a model and to estimate this functional relation between the extraneous variable and the dependent variable from the data. With this estimate one tries to extract the extraneous variable from the single measurements and to perform a convenient statistical analysis, e.g., an analysis of variance with correspondingly adjusted values, which are called residues.

In experimental designs with a pretest or a baseline (cf. Section 4.12), the subject is often regarded as a disturbance factor as far that the independent variable has only a rather small effect which cannot be detected due to the large variation of the initial values of the subjects. In such cases, one often assumes that the observed measurement is the sum of an individual value, which is constant for a fixed subject and a treatment value, which is different for different treatments. By calculating the difference between posttest and pretest one attempts to extract the individual value such that the treatment value alone can be obtained. The residue, which in this case is a **difference score**, is also denoted a **gain score**. If not only one pretest value but a baseline of several pretest values is available, the mean of the baseline values is often subtracted from the posttest score.

In another model, which is often used if the dependent variable is a response frequency, one assumes that the observed measurement is the product of an individual value and a treatment value. Here, the ratio of posttest and pretest is used in order to extract the individual value to obtain the treatment value. Instead of the ratio, mostly the equivalent **percentage change** is used where the difference score is divided by the pretest value and the result is multiplied by 100.

As one cannot assume that treatment effects are independent of the size of the initial values, i.e. are identical for all subjects of a treatment group, there is no way to decide empirically which of the two models, the additive or the multiplicative, is the true one. One even cannot assume that the relation between pretest and posttest can be expressed by a simple addition or multiplication. Therefore, both kinds of adjustments may produce completely useless residues, and an analysis performed with these residues may yield results, which can be trusted far less than the results of the same analysis, if it is applied to the unadjusted posttest scores.

If the score of an extraneous variable is available for each subject in addition to the score of the dependent variable, an analysis of covariance is often performed. In this kind of analysis the scores of the dependent variable are predicted by the scores of the extraneous variable, which here is denoted as **covariate**, by means of a linear regression. The residues are calculated by subtracting the predicted value from each value of the dependent variable. With the residues an analysis of variance is performed. However, when performing the procedure one has to consider in addition that the estimate of the regression line is distorted by the prediction error.

The analysis of covariance, as it is performed in most cases, assumes the validity of several assumptions, which are not very plausible. First, the covariate must be measured without error. This may approximately be true for extraneous variables as age or gender. However, this assumption cannot be valid for a pretest which was obtained with the same measuring instrument as the posttest for which one assumes that it is distorted by a measurement error. Second, it is difficult to believe that the relation between the covariate and the dependent variable should be linear. Third, one requires the regression lines to be parallel for all treatments which means that the influence of the treatments on the relationship between the covariate and the dependent variable should be the same for all treatments.

In general, one has to conclude from the above that a statistical adjustment for controlling extraneous variables had better be avoided in most cases, as the possibility for drawing causal conclusions is not necessarily improved in view of the required not very plausible assumptions.

Summary

1. Causal conclusions based on the outcomes of experiments are only possible if alternative explanations, i.e. threats to validity, are made implausible. Alternative explanations result, if causal variables, which are not to be studied, influence the used effect variables or might have such an influence. Such unwanted effect variables are called extraneous variables.

2. One of the most important techniques for controlling extraneous variables is randomization, i.e. the random assignment of subjects to experimental conditions or the random assignment of experimental conditions to subjects, respectively. In contrast to all other techniques, a randomization admits to control all potentially infinitely many known or unknown extraneous variables.

3. A further important control technique is the use of control groups, which comprise, e.g., placebo groups, yoked control groups, expectancy control groups, control groups for the effect of pretests, comparison groups and historical controls.

4. Global techniques for controlling extraneous variables are, e.g., elimination, blocking off, constancy, covering, balancing, counterbalancing, and blinding.

5. Local techniques for controlling extraneous variables are matching or blocking, respectively, which correspond to local constancy.

6. The usefulness of some common control techniques, as counterbalancing, repeated measures or the use of statistical adjustment must be doubted as well as the usefulness of matching or blocking.

Questions

4.1. Give a presentation as a two-factor design for the experimental design in Figure 4.4 similar to those in Figure 4.1 and 4.2 and explain the difference to the design in Figure 4.3.

4.2. Give an example, where the different timely order of two factors may yield different effects. Consider here the discussion in Section 4.1.4.

4.3. Explain the shortcomings of the design in Figure 4.5 in Section 4.1.4.

4.4. How many level combinations must be considered if Figure 4.6 in Section 4.1.4 is extended to three phases?

4.5. Which selection effects can occur if the proceeding of Riecken et al. (1974) with respect to the choice of the point of time of randomization is used as it was described in Section 4.1.5?

4.6. Give an applied example to illustrate why the study design of Zelen (1979) should be preferred to the approach of Riecken et al. (1974).

4.7. Choose in the randomized-play-the-winner rule $u = 3$, $v = 4$ and $w = 2$, and assume that successively four patients are to be assigned to therapy A or B and that the feedback about the success of a therapy is always known before the next patient enters the study.

Further, assume that patient 1 corresponds to case 1, patient 2 to case 4, patient 3 to case 4, and patient 4 to case 2 in the presentation of Section 4.1.5. Declare, how the composition of the cards changes after each feedback and how correspondingly the probabilities for assigning the two therapies change.

4.8. Which disadvantages may be present if the control technique of covering is used?

4.9. If the control technique of constancy is used, in which way should the necessary information be transmitted in case of human subjects or animals, respectively?

4.10. Which advantages and disadvantages have matching and blocking?

4.11. What would have been the consequence in the example for over-matching in Section 4.4, if a randomization after matching would have been possible?

4.12. 84 male students participated in a study (cf. [1]) on the effects of active and passive coping (coping is the way in which subjects respond to problematic situations). They were presented addition tasks via the screen of a computer for 20 minutes. In each task two double-digit integers were presented which had to be added mentally and the result had to be typed into the computer. A first independent variable was the difficulty of the tasks. This factor was varied in two levels. For the first level the subjects had to type in the first digit within 3.3 s, for the second level within 6.1 s. The time between two tasks was held constant so that the subjects had to perform 172 tasks for the first level and 123 tasks for the second. If an answer was wrong or came too late, a "–" appeared on the screen for 1 s, otherwise a "+".

A second independent variable concerned the control of the subjects by an aversive tone with a duration of 1 s. For the first level of this factor this aversive tone was given after each wrong or late answer. For the second level no aversive tone was given. For the third level the aversive tone was given independently of the performance of the corresponding subject if a partner which was assigned to this subject and to which the first level of the factor was applied received an aversive tone.

As 2 × 3 = 6 level combinations result for two factors with two or three levels, respectively, the total sample of 84 subjects was divided into 6 subsamples each with 14 subjects, which were assigned to the 6 level combinations. Four physiological measures and the responses to a questionnaire served as dependent variables.

a) Which information is missing which is necessary to decide whether causal conclusions are possible for this design?
b) Is the criticism of Church (1964) which was mentioned in Section 4.10.4 also valid for the present design with respect to the second factor and how this criticism had to be formulated here?
c) Which alternative explanations for possible outcomes of the experiment result in view of the fact that in each of three of the six groups 172 tasks had to be solved, however, in each of the three remaining groups only 123 tasks?

4.13. Discuss which shortcomings are present in the study performed by King Psammetichos II which is described at the end of Section 4.11. In which way could this study be improved?

4.14. Discuss the advantages and disadvantages of the use of repeated measures as a control technique.

5 Preliminary Experiments and Pilot Studies

Preliminary experiments are used to check whether it is possible, in principle, to perform a planned study. In preliminary experiments one tries to find out, for individual subjects, whether it is possible to record the dependent variables as planned and with the necessary reliability. This reveals whether the measuring instruments, e.g. the devices for measuring physiological parameters or questionnaires or observers, work without any problem and yield data of the required quality. Here, it is important to identify and remove any source of disturbance. In case of automatic data recording one tries to check the correctness of the values, e.g., by means of a manual control.

Further, one has to check whether the levels of the independent variables were realized as planned. Often the preliminary experiment already shows that certain levels have no effect on the dependent variable or, that the effects of different levels do not differ. These outcomes can serve as a basis to decide if the originally scheduled levels should be replaced by others.

In many cases, the preliminary experiments might be already quite time consuming, which has to be taken into account in advance, as the necessary corrections will usually yield a satisfactory result only after repeated modifications of the procedure as it was originally planned. Quite often the dependent and independent variables as they were originally scheduled turn out to be not suitable for the problem in question and have to be replaced by proper alternative variables.

After the preliminary experiments one should perform a **pilot study** or **pilot experiment** before starting the main experiment, i.e. a kind of dress rehearsal for the main study, which itself is usually quite expensive with respect to costs and time. In contrast to the main study considerably fewer subjects are used in the pilot study, e.g., only two subjects for each level or level combination, respectively, of the independent variable(s). If two factors are being investigated, each with two levels, i.e. four groups, one would need, e.g., 8 subjects in all for a pilot study. A pilot study should be performed with the same experimenters, the same instruments, in the same laboratories, under the same environmental conditions with respect to light, temperature, noise, etc. and exactly with the same time schedule as the planned main study. The pilot study and the main study should only differ in sample size. It is important to record the exact time needed for a single trial in order to be able to forecast the total time required for the main study. If the pilot study can be performed without a problem, the main study may be begun. However, in many cases further sources of disturbance might be identified as well as shortcomings in the operationalizations of the dependent and independent variables which were not detected during the preliminary experiments. Here, we have a last opportunity for corrections.

Some researchers tend to overinterpret the results of preliminary experiments and pilot studies, with the effect that hypotheses which have been formulated in advance on the basis of theories and former studies by these and other researchers are altered before the main study is being started. In some extreme cases, a directed hypothesis (e.g., "Treatment A is more efficient than treatment B.") is altered into the reverse hypothesis (i.e. "Treatment B is more efficient than treatment A."). However, in most cases a more specific hypothesis will result, where an undirected hypothesis (e.g., "Treatment A differs from treatment B.") is reformulated into a directed hypothesis (e.g., "Treatment A is more efficient than treatment B."). Though a researcher is always free to formulate whichever hypotheses he or she likes, as long as this is done before the study is performed, note that the results of pilot studies, due to the small sample sizes, are in general without any evidence for the formulated hypotheses. This explains why the outcomes of pilot studies are often the exact opposites of the results of the main study.

Hence an argument becomes questionable which is often used as a justification for the performance of pilot studies. According to this argument, pilot studies are being performed in order to make a prediction of whether it is worthwhile to perform the main study.

In spite of the small sample sizes, pilot studies often require a considerable amount of effort and costs. Therefore, sometimes the outcomes of a pilot study are added to those of the main study, thus reducing the effort needed for the real main study. Such a proceeding must be categorically rejected, because one can never rule out the possibility that both studies differ systematically in some respect.

Summary

1. In preliminary experiments one checks for single subjects whether the levels of the independent variable were selected such that effects can be detected and whether the reliability of the dependent variable is sufficiently high.
2. After the completion of the preliminary experiments a pilot study based on few subjects is performed in order to check whether the main study can be performed without problems. One must not add the outcomes of the pilot study to those of the main study.

Questions

5.1. Is it possible to dispense with a pilot study, if enough preliminary experiments were performed?

5.2. What might happen, if the outcomes of a pilot study are added to those of the main study?

5.3. Is it necessary to perform a further pilot study before starting the main study, if, due to the outcome of a pilot study, essential modifications are necessary?

6 Designs which Had better be Avoided

6.1 Designs without Randomization

If the subjects in a study are not randomly assigned to the levels or level combinations of the independent variable(s), so-called nonequivalent groups or a quasi-experiment result. As discussed in Section 4.1.2, causal conclusions, in principle, are impossible in these situations, since one cannot rule out that subjects from different groups differ in characteristics which have an effect on the dependent variable. One might try, at most, in view of the potentially infinitely many possible alternative explanations for the observed outcomes, to make the groups as similar as possible with respect to all known or suspected interfering variables. One might try to rule out the most obvious alternative explanations by using control techniques, such as, e.g., matching or blocking (cf. Section 4.4). For this, it might also help to study as many appropriate comparison groups (cf. Section 4.10.7) as possible. It is always better, however, to put up with very small sample sizes, than to do without a randomization.

One reason for the use of designs without randomization in clinical studies is the experimenter's refusal to give no treatment or only an inefficient treatment to patients. In most cases, ethical reasons or legal arguments are brought forward, which do not permit that the best treatment available is withheld from patients, because of a random decision. Thus, the starting point is the assumption that the respective doctor can determine the optimally efficient therapy for a patient after an appropriate diagnosis. If this assumption is actually true, the corresponding clinical trial and, thereby, also any randomization in such a trial is indeed superfluous and the study should not be performed.

If one is, however, not sure of the efficacy of a treatment, the ethical argument above should be refuted, because the use of a potentially not efficient treatment is ethically not more justified than the performance of no treatment or of a sham treatment. In view of the studies about the efficacy of placebos (cf. Section 4.10.2) in such a case a sham treatment without supposed side-effects most probably is ethically more appropriate than a treatment with a doubted efficiency with possible adverse effects.

In studies with human subjects or patients there may exist reasonable ethical reasons which exclude the use of a randomization. If, e.g., one wants to investigate whether smoking causes lung cancer, a prospective study might use a sample of 1000 pre-school children 500 of which are randomly selected. From their sixteenth year onward they have to smoke at least 10 cigarettes a day with a defined risk potential, while the remaining 500 children are not permitted to smoke all their life. As a dependent variable one records whether a person falls sick with lung cancer before the 50th year of life. Though this question is obviously very important and though the above reasonable design, in principle, can be realized, such a study cannot be performed for ethical and legal reasons. This means that comparison groups of smokers and non-smokers are studied for which it is unknown, in spite of a matching with respect to many known variables like age, gender etc., whether they differ in a characteristic which yields a high predisposition for lung cancer without smoking being causally responsible for lung cancer. E.g., a gene might exist that has a certain importance for addictive behavior as well as simultaneously for falling sick with lung cancer.

Another example is the question whether stress caused by a certain academic examination has an effect on the immune system of the corresponding students. Theoretically, it is possible to split up a sample of students at random into two halves, where the first half has to pass the examination in exactly six months, while the other half has to pass it in exactly two years. Such a procedure with an arbitrary chosen delay of an examination is technically possible, but legally not permitted. This means that a matched comparison group of students without the respective examination has to be found to a sample of candidates who are going to pass this examination and who have agreed to participate in the study. The outcome of the study has to be interpreted with caution since selection effects cannot be ruled out.

As a consequence of this reasoning one should always try to randomize even if very small sample sizes may result or the original problem has to be reformulated. If a randomization is not possible for a given problem for ethical or legal reasons, but if at the same time an answer to the prevailing question is very important, one can try to make the most obvious alternative explanations (threats to internal validity) implausible by using a series of suited comparison groups and possibly by using additional measures. When using this approach one cannot be sure, however, whether any detected effects do actually exist.

6.2 Designs without a Control Group

As discussed in Section 4.10.1, in addition to an appropriate randomization the existence of one or more suited control groups is necessary in most cases if one wants to be able to draw causal conclusions. If only one treatment is applied and a dependent variable is recorded afterwards one cannot conclude that the treatment had any effect. This is true even if first a pretest is performed, then a treatment is applied and finally a posttest is performed. Any differences between pretest and posttest are not necessarily due to the treatment as discussed in Section 3.2.1, 3.2.2, 3.2.3, and 3.2.4.

If instead of a one-group design with one treatment, a one-group design is used in which in a certain timely order treatment and control conditions are provided, a within-subjects design (cf. Section 4.12) results, in which problems with respect to the interpretation of the outcomes are inherent even with counterbalancing (cf. Section 4.8). Causal conclusions cannot be made if instead of a true control group historical control groups are used (cf. Section 4.10.8). As the discussion of the cockroaches example (cf. Section 4.11) by Matheson, Bruce, and Beauchamp (1971) shows, the necessity of control groups also exists for a conservative arrangement of the levels of the extraneous variables. In the study performed by King Psammetichos II (cf. Section 4.11) it would have also been reasonable to consider, apart from the sample of boys which were brought up together with goats, another independent sample of boys who were brought up by women whose tongues had been cut off, as well as a third independent sample of boys who were brought up together with goats whose tongues were cut off.

The example above demonstrates very nicely that an appropriate experimental design may depend on the cultural epoch in which a study is being performed. While the Hellene Herodotos rejected the idea of bringing up boys by women with cut-off tongues as "foolish", it does not seem unimaginable that during that epoch cultures existed, e.g., in Egypt, where cutting-off of tongues of women was considered as a common and advisable usage. Similarly we cannot rule out that many of the experimental designs for laboratory animals which are accepted in our culture by many people, might no longer be permitted in 50 years. Nevertheless, let's come back to Herodotos: the Hellenic culture did obviously not disapprove of taking away infants from their mothers in order to bring them up with goats!

In studies with two or more treatments applied to different groups, where no treatment corresponds to a control, causal conclusions are possible. However, we saw in Section 4.10.3 that even in such cases the interpretation of the outcomes can be improved considerably by including control groups.

Altogether, one should always wonder in designs without a control group, whether the design is complete, i.e. whether it is possible on the whole to draw causal conclusions without such a group. If patients are assigned to two therapy groups using randomization because it is not permitted for ethical or legal, respectively, reasons to use a sham treatment or no treatment at all (see the case of the study with contraceptives in Section 4.10.2!) the absence of a control group has the consequence that, depending on the outcomes of the study, one does not know whether both therapies, one therapy or no therapy has had a positive effect. Here, one should consider using a standard therapy with established and known effects as one of the two therapies or using such a therapy in an additional control group. This would considerably improve the interpretability of the outcomes. Nevertheless, the judgement of the standard therapy strictly speaking would be based on historical controls (cf. Section 4.10.8) with known disadvantages.

6.3 Designs with Repeated Measures

In Section 4.1.4, 4.8, 4.10.6, and 4.12 we discussed in detail the problems which arise when the outcomes of designs are interpreted in which repeated measures are provided. In particular, this concerned within-subjects designs, in which all levels of the independent variable staggered as to time are assigned to a single sample. Because of the difficulties which arise when designs with repeated measures are being used, such designs should be avoided if one intends

1. to get designs with a higher precision, i.e. designs where the error variance is smaller than for other designs,
2. to get designs where the number of required subjects is smaller than for between-subjects designs.

The first intention is based on the idea that in designs with repeated measures each subject might serve as its own control. The underlying assumption is that each measurement is the sum of an "individual value" and a "treatment value". Assuming this, we get a so-called **gain score** which is a measure of the pure treatment effect, by subtracting scores recorded under the control condition from scores recorded under the treatment condition. The individual values cannot be separated from the treatment values, however, if a between-subjects design is being used with independent samples for treatment and control condition. In this case, a considerable part of the variance of the measurements is the result of the differences between the subjects. A large variance of the measurements prevents small differences between the treatments from being detected. Because of the assumption that it is possible to separate the individual values from the treatment values by computing differences when within-subjects designs are being used, the variance of the measurements is reduced to the variance of the treatment values and the chance that the existence of treatment differences can be proved is increased. Therefore, it is often formulated that a within-subjects design has a higher **precision** than the respective between-subjects design.

A problem with this kind of argumentation is that it makes the unplausible assumption that it is possible to separate the individual value from the pure treatment value, simply by computing a difference of scores. In fact, we have seen that one can never, in the case of several measurements for each subject, rule out that pretests have effects on posttests and that measurements are not necessarily determined by individual values and treatment values alone, but may also be influenced, e.g., by history or maturation. If more than one treatment is applied to a subject, one has to consider that measurements might not only be composed of individual and treatment values, but also, e.g., of transfer effects which are due to preceding measurements and/or treatments.

For these reasons one should use independent groups in repeated-measures designs in order to control the effects of several measurements for each subject (cf. Section 4.10.6) as well as to separate transfer effects from pure treatment effects (cf., e.g., Section 4.1.4). The use of many groups of subjects causes a considerable increase in the necessary sample size by comparison with a corresponding between-groups design.

As neither of the two objects mentioned above, which are to be achieved by repeated measures can be realized, it is the best, as a rule, to dispense with such designs. An exception can only be made, if the effect of multiple measurements or multiple treatments, respectively, itself is the object of investigation. In these cases it is necessary to provide for corresponding control groups for multiple measurements or multiple treatments, respectively, to be able to interpret the resulting outcomes appropriately. This has the consequence that considerably larger sample sizes are required as for the convenient between-groups designs.

6.4 Crossover Designs

Crossover designs are a mixture of within- and between-groups designs and have already been discussed in the context of the control technique of counterbalancing (cf. Section 4.8). The most simple and probably the most frequently used **crossover design** is used for the comparison of two treatments A and B, where one of the treatments might also be a sham treatment, thus, corresponding to a control condition. In this design (cf. Figure 6.1) a sample of subjects is randomly split into two subsamples each having half the size of the original sample. In a phase 1, one subsample gets the treatment A, the other subsample the treatment B. After the treatments, a measurement is recorded in both samples. If these measurements were consequently used for a comparison of the two treatments, this would yield a simple between-groups design and a causal interpretation would be possible, in principle. In the crossover design, however, the two treatments are given again in a phase 2, this time in an inverted assignment. Thus, we now have two groups, one of which first gets treatment A and then treatment B (group A_1B_2), while the other group first gets treatment B and then treatment A (group B_1A_2), as depicted in Figure 6.1. After the treatments have been applied in phase 2, measurements are recorded again.

		Phase 2	
		Treatment A	Treatment B
Phase 1	Treatment A		A_1B_2
	Treatment B	B_1A_2	

Figure 6.1: Simple crossover design with two treatments A and B

The following advantages are ascribed to this design:

1. By pooling the posttest values for phase 1 and phase 2 for each treatment (A and B) the number of measurements is doubled for each treatment, as opposed to the design with the two groups A_1 and B_1. Thus the chance for detecting a treatment difference should be increased. In other words: one hopes to be able to make a comparison, in this way, which is as efficient as a simple two-group design with twice the sample size. By counterbalancing the timely order of the treatments one wants to attain that possible effects cannot be attributed to the fact that one particular treatment is always being given as the first and the other one as the second. Otherwise, the immediate alternative explanation that any effects may be due to the fixed order of treatments would be at hand.

2. In particular, in clinical studies each patient receives an effective treatment in this design. Hence ethical arguments against a no-treatment control group are no longer applicable.

Considering the discussions in Section 4.1.4, 4.8, 4.10.6, and 6.3 it is immediately obvious that the first apparent advantage of a crossover design exists only if completely implausible assumptions hold and this design will yield outcomes which cannot be causally interpreted in usual realistic situations: by recording two measurements for each subject, effects of the first measurement on the second measurement cannot be ruled out, where, in addition, the presumable dependence of the two measurements will cause problems in the statistical evaluation. As each subject receives two treatments in a timely order, effects of the treatment of phase 1 on the measurement after the treatment in phase 2 cannot be ruled out. Such outlasting effects of treatments are called **carry-over effects** or **transfer effects**. If an outlasting of treatment effects yields difficulties when the effects of succeeding treatments are estimated, the term carry-over effect is used. But if only the after-effects of treatments on succeeding treatments are of interest the term transfer effects is used.

Carry-over effects can result in an additive accumulation of the different treatment effects as well as in an alteration of the effect of a following treatment which is caused by a preceding treatment. In particular, a preceding treatment can have the effect that the effect of a following treatment is prevented.

A pooling of measurements from both groups after treatment A or B, respectively, can have the effect that each possible result is feigned: a non-existing superiority of treatment A or, alternatively, of treatment B may be "detected". But it is also possible that the detection of a really existing superiority of one treatment is prevented by the influence of carry-over effects. As a rule, the outcomes of a crossover design as depicted in Figure 6.1 cannot be interpreted conclusively.

One possibility to derive a causal conclusion for the outcomes of such a design consists in considering only the measurements of phase 1 and comparing A_1 with B_1. But then, the treatments B_2 and A_2 together with the corresponding measurements become superfluous and only cause additional stress for the participating human subjects or animals. Depending on which kind of treatment or measurement is being used, it might also be that such a crossover design must be refused for ethical reasons.

The second advantage of the crossover design, mentioned above in connection with clinical studies, must also be doubted. If both treatments are effective, the effect of the treatment in phase 2 cannot only suppress the effect of the treatment in phase 1, but can even result in a harmful effect due to an interaction of the two treatments. If treatment A, however, is a sham treatment and only treatment B is potentially effective, and if both groups are to definitely receive treatment B, one should apply the combination A_1B_2 without measuring after B_2 to the first group and treatment B_1 alone with the following measurement to the second group. Then, a causal conclusion can be drawn by comparing the measurements after A_1 or B_1, respectively, and both groups have received, as required, treatment B. The first group would be saved the measurement after B_2 and the second group would be saved the superfluous treatment A_2 with the subsequent measurement.

The lack of interpretability of the crossover design is already indicated in Figure 6.1 as the two empty cells seem to call for missing groups. By completing Figure 6.1 to Figure 6.2 a design results with outcomes which can be readily interpreted.

		Phase 2	
		Treatment A	Treatment B
Phase 1	Treatment A	A_1A_2	A_1B_2
	Treatment B	B_1A_2	B_1B_2

Figure 6.2: Completed crossover design for two treatments A and B

An initial sample of subjects is randomly divided into four subsamples which correspond to the treatment combinations A_1A_2, A_1B_2, B_1A_2, and B_1B_2. If the measurements after treatment A_1 in the groups A_1A_2 and A_1B_2 are pooled and likewise the measurements after treatment B_1 in the groups B_1A_2 and B_1B_2, the comparison of these now two groups of measurements shows whether the two treatments A and B have different effects.

However, if one compares the measurements after treatment A_2 in the groups A_1A_2 and B_1A_2 or the measurements after treatment B_2 in the groups A_1B_2 and B_1B_2, respectively, one can investigate the strength of the effects of the two initial treatments on the outcome of the corresponding second treatment. A similar comparison for the groups A_1A_2 and A_1B_2 or B_1A_2 and B_1B_2, respectively, reveals the strength of the effect of an initial treatment on the differences between the second treatments.

The difficulties in interpreting the outcomes of the crossover design with two treatments and two periods (two-treatment-two-period crossover design or two-period changeover design) in Figure 6.1 have been known for a long time, and it is difficult to understand, why this design, which nearly never yields outcomes which can be conclusively interpreted is still used in many studies. Actually, already Grizzle (1965) pointed out that this design can only be used, if two assumptions are made, which are very difficult to justify, namely, first that carry-over effects (transfer effects, **residual effects**) are identical for both treatments, and second that the correlation of both measurements is positive for each subject. Brown (1980) and Freeman (1989) demonstrated that the two-stage evaluation procedure for this design which was proposed by Grizzle (1965) may be misleading. Fleiss (1989) points out that many attempts to improve this design or to improve corresponding evaluation methods result only in new problems and that this design may only be used, if the state of the subjects does not change during the duration of the study and if possible transfer effects are the same for both treatments.

One tries to avoid transfer effects by using long **wash-out periods** between the treatments which, however, has almost inevitably the consequence that the requirement of a stable state of the subjects which was required by Fleiss (1989) cannot be met. Furthermore, usually one cannot expect that patients accept these long treatment interruptions in clinical trials. Though such patients may continue to participate in the study a part of them will try to get help outside the study at the same time, such that the wash-out periods are no longer effective. On the other hand, with treatment interruptions, which are too short, the probability of transfer effects is increased. Actually, transfer effects also cannot be avoided with long interruptions, if the treatments cause irreversible effects. It is not advisable to make the duration of a wash-out period conditional on whether traces of the drug can still be found in the body after this period or not. The duration of the period should rather be long enough to ensure that also no behavioral effects of the drug remain, though this may be difficult to ascertain.

Laska, Meisner and Kushner (1983) proved that the design which is depicted in Figure 6.2 is universally optimal, i.e. that this design fullfils all of three well-known mathematically defined optimality criteria for experimental designs (D-, A- and E-optimality). This is also indicated by Laska and Meisner (1985), Carrière and Reinsel (1992), and Carrière (1994). Nevertheless, problems arise also for this design as for its realization and the interpretability of the outcomes.

One difficulty with this design concerns the problem of introducing the treatment combinations A_1A_2 and B_1B_2 in such a way that A_2 and B_2 are really comparable with the second treatments in the combinations A_1B_2 and B_1A_2. If, e.g., treatment A is a sham treatment, i.e. if A corresponds to a control condition, it may be difficult to apply the control condition for the combination A_1A_2 two times in a row. The two treatments cannot be realized as being separate treatments by the subjects, as opposed to the combinations A_1B_2 and B_1A_2. In such a situation the combinations A_1A_2 and A_1B_2 do not only differ in the second treatment, while the first treatment is the same, but also in the perception of the second treatment. This inability to perceive two equal treatments as being separate which are only separated in time by a formal definition, may also exist, of course, if A (or B, respectively) is not a sham treatment.

Of course, these weaknesses of the optimal design which is depicted in Figure 6.2 have been known for a long time and quite a few authors have made proposals to improve this design. In order to systematize the different kinds of crossover designs, such designs are generally denoted designs of the type COD (t, p, s). Here, COD stands for "crossover design" or "change-over design", t for the number of treatments, p for the number of periods and s for the number of sequences, i.e. also for the number of independent samples. The design in Figure 6.1 is of the type COD (2, 2, 2), while the design in Figure 6.2 is of the type COD (2, 2, 4).

The disadvantages of the quite common crossover design of the type COD (2, 2, 2) have been known for a long time and have also been discussed by us. Proposals to improve this design are usually related to the use of pretests, of more periods or of more sequences. A survey of several such approaches is given by Carrière (1994). This author considers, in particular, designs with p = 3 periods and shows that the interpretability of the outcomes of different designs depends on which assumptions are made with respect to the different kinds of carry-over effects. According to the analyses of this author the design of the type COD (2, 3, 4) with the sequences $A_1B_2B_3$, $B_1A_2A_3$, $A_1A_2B_3$, and $B_1B_2A_3$ seems to yield outcomes which permit conclusive interpretations, at least if the validity of one of the four models considered by the author is assumed. Note that the above design corresponds to the optimal design in Figure 6.2 where a further period is added to each sequence. A corresponding four-period design of the type COD (2, 4, 4) with the sequences $A_1B_2B_3A_4$, $B_1A_2A_3B_4$, $A_1A_2B_3B_4$, and $B_1B_2A_3A_4$ was discussed by Matthews (1990). A disadvantage of all of these designs with more than two periods consists, without doubt, in the possibility that now not only carry-over effects on the subsequent treatment but also on later treatments are possible.

6.5 Designs with more than Two Factors

At first it seems to be advisable to consider the effects of as many independent variables, i.e. factors, as possible on one or more dependent variables. With this method, it seems possible to detect not only one but several and even more complicated potential causal relations at the same time. Additionally, extraneous variables can possibly be controlled by considering them as independent variables as discussed in Section 4.5. The more factors are considered at the same time, the larger the amount of observed variance of the dependent variables which can be explained by these factors, i.e. the smaller the amount of unexplained variance (**error variance**) which is left. Theoretically, one can conceive that all factors which influence the dependent variable are considered, such that it is possible to predict exactly each value of the dependent variable, if the corresponding levels of the factors are known. In practice of course, one generally does not know all relevant factors and even if one knew them, one could not, due to the large number of factors, take all of them into account.

As already discussed at more than one point, e.g. in Section 6.3, we cannot recommend the usage of **repeated-measures factors** where the factor levels correspond to the level of an independent variable at different points of time because this yields outcomes which cannot be easily interpreted. But if the factor levels do not correspond to such repeated measures, an independent sample of subjects must be available for each combination of factor levels. This considerably augments the necessary sample size if one is only interested in effects of the single factors, the so-called **main effects**.

If one is interested, e.g., to know whether four different treatments have an effect on a dependent variable, one can consider four factors each with two levels which correspond either to the respective treatment or to a respective control condition. To find out which of the four factors have an influence on a dependent variable, 80 subjects are randomly allocated to eight subsamples each consisting of 10 subjects, and each subsample is randomly assigned to one of the four treatment or four control groups (cf. Figure 6.3). A comparison is performed for each of these four independent one-factor designs, with two factor levels each.

Factor 1		Factor 2		Factor 3		Factor 4	
T1	C1	T2	C2	T3	C3	T4	C4

Figure 6.3: Experimental design used in order to find out, which of the four treatments T1, T2, T3, and T4 in comparison with the respective control condition C1, C2, C3, or C4, respectively, has an effect on the dependent variable

If the same control condition can be used for all four treatments, the total sample size is reduced to 50 subjects which are randomly divided into five subsamples each with 10 subjects, which are randomly assigned to the four treatments and the control condition (cf. Figure 6.4). Here, we have a one-factor design with a factor which has five levels. Each treatment group is compared with the same control group.

Factor				
C	T1	T2	T3	T4

Figure 6.4: Experimental design used in order to find out which of the four treatments T1, T2, T3, and T4 has an effect on the dependent variable in comparison with the control condition C

A considerably larger sample size is required for the four-factor design in Figure 6.5. Here, 160 subjects have to be randomly divided into 16 subsamples with 10 subjects each and each subsample has to be randomly assigned to one of the 16 possible combinations of factor levels. In Figure 6.5, e.g., the combination of factor levels No. 7 corresponds to the combination where treatment is effective for factor 1 and factor 4 at the same time, while the control condition is effective for factor 2 and factor 3.

We already discussed in Section 4.1.4 that it is not always possible that several factors are effective at the same time, and proposals were made to overcome the resulting difficulties in interpreting the outcomes by more complicated designs (cf. Figure 4.6) with even more independent groups. This problem can also have the consequence that one confines oneself to a one-factor design. This solution is also suggested by the problem of the large sample sizes which are needed in multifactorial designs, as already discussed.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T1 T2 T3 T4	T1 T2 T3 C4	T1 T2 C3 T4	T1 T2 C3 C4	T1 C2 T3 T4	T1 C2 T3 C4	T1 C2 C3 T4	T1 C2 C3 C4	C1 T2 T3 T4	C1 T2 T3 C4	C1 T2 C3 T4	C1 T2 C3 C4	C1 C2 T3 T4	C1 C2 T3 C4	C1 C2 C3 T4	C1 C2 C3 C4

Figure 6.5: Listing of the 16 possible combinations of factor levels for four factors, each with one treatment level (T) and one control level (C)

An undeniable advantage of multifactorial designs is the possibility to detect differential effects which cause so-called interactions. If one is less interested in the pure effect of factors, which are also called **main effects**, but would like to find out which specific effects a certain factor has in the presence of given level combinations of other factors, a multifactorial design has to be used, since this question cannot be answered by the usage of a one-factor design as depicted in Figure 6.3 and 6.4. If the effect of a factor on a dependent variable is always the same, irrespective of which level combination of the other factors is effective at the same time, one says that the corresponding factor does not interact with the other factors, or that no **interaction** of this factor with the other factors exists.

In order to illustrate the concept of interaction, we give fictitious population means for a design with two factors–each with two levels–in Figure 6.6. As a simplification, we assume that 10 subjects were assigned to each of the four level combinations, that all 40 subjects were completely equal before the experiment, and that the measurement of the dependent variable is performed for all subjects without any errors. In this case all 10 subjects in one group would have exactly the same score and, of course, this score would be equal to the mean of the 10 scores. This value would be entirely determined by those levels of the two factors which affect the subject of the corresponding group.

	C2	T2
C1	10	20
T1	15	30

Figure 6.6: Population means of the dependent variable for a design with two factors, each with the levels treatment (T) and control (C) for illustrating differential effects in the same direction

Figure 6.6 reveals that the difference between treatment T1 and the corresponding control C1 is given by 15 – 10 = 5, if the control condition C2 of the second factor is effective at the same time. If, however, the treatment condition T2 of the second factor is effective, the difference between treatment T1 and control C1 is given by the value 30 – 20 = 10, which is twice as large. The extent of the effect of treatment T1 as opposed to control C1, here, depends essentially on the level of the second factor which is simultaneously effective. Whenever the extent and/or the direction of an effect, which is caused by a factor with respect to a dependent variable, depends on the levels of one or more other factors, which are effective at the same time, a so-called **interaction** exists. In the present case, it is not sensible to ascribe a certain effect to treatment T1 as the extent of this effect depends on the level of the second factor. Similar **differential effects** can also be observed, by the way, for the second factor in Figure 6.6, if the levels of the first factor are kept constant. Under C1 the difference between T2 and C2 is given by 20 – 10 = 10, while under T1 this difference amounts to 30 –15 = 15.

	C2	T2
C1	10	20
T1	10	30

Figure 6.7: Population means of the dependent variable for a design with two factors, each with the levels treatment (T) and control (C) in order to illustrate no effect of the first factor under C2 and a positive effect of the first factor under T2

In Figure 6.7 no difference between T1 and C2 is found if level C2 of the second factor is fixed (because of 10 – 10 = 0). However, a positive difference results if level T2 is fixed (because of 30 – 20 = 10). Again, we have an interaction.

	C2	T2
C1	10	20
T1	15	5

Figure 6.8: Population means of the dependent variable for a design with two factors, each with the levels treatment (T) and control (C) in order to illustrate opposite differential effects of the first factor for the two levels of the second factor

In Figure 6.8 a positive difference between T1 and C1 is found, if the level C2 of the second factor is fixed (because of 15 – 10 = 5 > 0), but a negative difference, if the level T2 is fixed (as 5 – 20 = –15 < 0). Again, we obtain different differential effects and, as a consequence, an interaction of the two factors.

	C2	T2
C1	10	20
T1	15	25

Figure 6.9: Population means of the dependent variable for a design with two factors, each with the levels treatment (T) and control (C) to illustrate differential effects of the same extent and the same direction of one factor for both levels of the other factor, i.e. for the case that no interaction of both factors exists

Finally, in Figure 6.9, we find the same effect of T1 relative to C1 with respect to extent and direction under C2 (because of 15 – 10 = 5) as well as under T2 (because of 25 – 20 = 5), i.e. the differential effects are the same, and no interaction exists. In the same way, we find the same difference for the second factor under both levels of the first factor (because of 20 – 10 = 10 and 25 – 15 = 10). This also holds in general, as if a factor does not interact with a second factor, the second factor also does not interact with the first factor.

Only in the case depicted in Figure 6.9 it is appropriate to assume a main effect (of size 5) for the first factor and a main effect (of size 10) for the second factor, because these effects are the same, irrespective of which level of the respective other factor is present. Because of this, these main effects may be interpreted in substance. This does not prove true if an interaction of the factors is present, as shown in the examples in Figure 6.6, 6.7, and 6.8. The statistical evaluation of a factorial design is usually performed by an analysis of variance. In this kind of analysis no consideration is given to the problem that main effects cannot be interpreted in the presence of interactions. In this kind of procedure the main effect of a factor is rather estimated by averaging with respect to the levels of all other factors. E.g., in Figure 6.8 the value (15 + 5) / 2 = 10 for T1 would be compared to the value (10 + 20) / 2 = 15 for C1 and a negative main effect of the extent 10 – 15 = –5 would be stated for the first factor, though–in reality–we have a positive effect of 5 for level C2 and a far more extreme negative effect of (–15) for level T2. Thus, if interactions are present, the results of an analysis of variance with respect to main effects cannot be easily interpreted.

The fact that there are tests for interactions in the analysis of variance is not a great help either. If such a test is significant we can conclude that the results of the test with respect to the main effects should not be interpreted, but we do not know what the differential effects of the factors are like. However, if a test for the existence of an interaction is not significant, we have seen in Section 3.1.1 that this does not mean at all that no interaction exists. Thus, the results of tests for main effects, again, should not be interpreted. As in the preceding examples, one cannot rule out that the direction of an effect is described correctly for certain level combinations of the other factors only and that information about the extent of a main effect is always misleading if interactions are present. In particular, in the example in Figure 6.10 an analysis of variance would not reveal a main effect though we always find a positive effect of the other factor under the control condition of one factor (because of 20 – 10 = 10), but, a negative effect under the treatment condition (because of 10 – 20 = –10).

	C2	T2
C1	10	20
T1	20	10

Figure 6.10: Population means of the dependent variable for a design with two factors, each with the levels treatment (T) and control (C) for illustrating effects of the same extent but of opposite directions for fixed levels of one factor

For an appropriate interpretation of the outcomes of a design with two factors, each with two levels, the four comparisons C1C2 vs. T1C2, C1T2 vs. T1T2, C1C2 vs. C1T2, and T1C2 vs. T1T2 should be performed, by which specific statements about the existence of differential effects could be made. This would, however, require four comparisons even in the most simple multifactorial design. If, as depicted in Figure 6.11, we had not only two but three levels for each of two factors, a total of 18 comparisons would be necessary if all possible differential effects were to be considered (A1B1 vs. A1B2, A1B1 vs. A1B3, A1B2 vs. A1B3, A2B1 vs. A2B2 etc.). A global interpretation will often be rather difficult in substance in view of so many possible differential effects.

	B1	B2	B3
A1	A1B1	A1B2	A1B3
A2	A2B1	A2B2	A2B3
A3	A3B1	A3B2	A3B3

Figure 6.11: Two-factor design with the levels A1, A2, and A3 of the first factor and the levels B1, B2, and B3 of the second factor

The problem that too many comparisons have to be performed in order to detect differential effects is still aggravated if not only two but three or even more factors are considered simultaneously. Figure 6.12 depicts the eight possible level combinations to which independent samples of subjects must be randomly assigned if three factors, each with two levels, are present. If one is interested, e.g., in differential effects of the first factor, group 1 is compared with group 5, group 2 with group 6, group 3 with group 7, and group 4 with group 8. Here, two groups which differ only with respect to the first factor are compared at each time.

1	2	3	4	5	6	7	8
T1	T1	T1	T1	C1	C1	C1	C1
T2	T2	C2	C2	T2	T2	C2	C2
T3	C3	T3	C3	T3	C3	T3	C3

Figure 6.12: Listing of the eight possible level combinations for three factors, each with a treatment level (T) and a control level (C)

In addition to the reasons given above (large sample size, it is not always possible to apply several factors at the same time), which explain why multifactorial designs had better be avoided, the number of possible differential effects may become that large that a simple general interpretation of the outcomes is no longer possible. Here, one might ask whether the interpretation of a one-factor design is actually more simple than that of a multifactorial design, as differential effects may exist irrespective of whether we use a design by which they can be detected or not. In other words: Are effects which are detected in one-factor designs nothing but artificial averaged main effects like those found by an analysis of variance?

As for independent variables which are manipulated by the experimenter this suspicion can be repudiated. If we are only interested in the first factor in Figure 6.12 and if we use a one-factor design, ignoring the second and third factor, this corresponds to the choice of two control conditions C^*2 and C^*3 with respect to these two factors. Thus, we compare only the two groups 4^* (corresponding to $T1C^*2C^*3$) and 8^* (corresponding to $C1C^*2C^*3$) instead of the eight groups, i.e. we study, in principle, a differential effect by fixing the level combination C^*2C^*3.

This argumentation holds also in a similar way for extraneous variables which are kept constant by the experimenter (cf. Section 4.3). The argumentation for the control of extraneous variables by means of randomization (cf. Section 4.1) is slightly more complicated. Here, in fact, "averaged" effects are considered and differential effects with respect to subpopulations of subjects cannot be determined. Nevertheless, this is in contrast to the lacking interpretability of "averaged" effects in a multifactorial design no problem of internal validity, but a problem of external validity, i.e. the causal conclusions drawn with respect to the effects of the considered independent variables are here justified. It is only an open question, for which subpopulations of subjects the detected causal conclusions are valid.

SUMMARY

1. If the assignment of subjects to the possible levels or level combinations of the considered independent variables is not performed in a true random manner, i.e. if no randomization is used, no causal conclusions can be drawn due to the large number of potential alternative explanations.

2. As it is, in general, not possible to draw a true random sample from a real population, one usually tries to derive causal conclusions by comparing the outcomes of designs with two or more groups. At least one of these groups should be a control group.

3. The use of repeated measures as well as the application of more than one treatment to the same subject lead to problems with respect to the interpretation of the outcomes because of possible after-effects. This holds, in particular, for crossover designs.

4. While the outcomes of one-factor designs are easy to interpret in most cases, this is no longer the case for designs with two or more factors, because the existence of interactions cannot be ruled out for the latter.

Questions

6.1. Give an example of how to improve the interpretability by additional measurements if no randomization is used.

6.2. Indicate when within-subjects designs should be used on no account and when they have to be used.

6.3. Why is the level of precision of an experimental design an unsuitable criterion for selecting a design?

6.4. Give a short argument, why the outcomes of the crossover design in Figure 6.1, in general, are not interpretable.

6.5. Explain, why the pooling of measurements after equal treatments can result in different kinds of misinterpretations for the crossover design in Figure 6.1.

6.6. Explain, why in a one-factor design the occurrence of "averaged" effects as a consequence of a randomization does not affect the drawing of causal conclusions, though causal conclusions with respect to "averaged" effects are not appropriate in multifactorial designs if interactions exist.

6.7. In a study (cf. [2]) 119 obese and 77 normal-weight females completed a questionnaire on gastrointestinal (GI) symptoms and on binge eating behavior. Binge eating means a cyclic disorder where during a restricted time interval a large quantity of high-caloric food is ingested secretly. In this study a subject was considered as obese, if her weight (measured in kg) divided by the square of her height (measured in m^2) exceeded 30. On the basis of this definition and on binge eating behavior the four subsamples "obese binge eaters" with 73 persons, "obese non-binge eaters" with 43 persons, "non-obese binge eaters" with 14 persons, and "normal-weight controls" with 61 persons were formed. It is reported, amongst other things, that "obese binge eaters" exhibited "significantly" more "upper GI symptoms" (e.g., nausea, flatulence, belching) than "normal-weight controls" or "obese non-binge eaters", and that "obese binge eaters" exhibited "significantly" more "lower GI symptoms" (e.g., abdominal pain) as "normal-weight controls". A certain number of further statistical results is reported some of which were "significant" and others were not.

Give reasons why the outcomes of this study do not permit any causal interpretation.

6.8. In a study investigating the defecation threshold in pigs (cf. [3]) three female Yucatan Micropigs and one female domestic pig were used. Each animal participated in four sessions, each lasting approximately 90 min, where a feeding period of about 10-15 min was included. The sessions were separated by one week. In each session an animal either received .05 mg/kg atropine or a placebo (saline). During the first two sessions a sequential order of the two conditions was used that was inverted during the next two sessions. The object of the study was to replicate former results about lowering the defecation threshold by eating and to investigate the effect of atropine on this reduction. The interpretation of the outcomes was that for each of the two dependent variables the scores for the altogether eight placebo conditions on the one hand and the scores for the altogether eight atropine conditions on the other hand were pooled, i.e. one acted as if scores of 16 animals had been recorded. Then, the eight placebo scores before eating were compared with the eight placebo scores after eating and the same was done with the eight atropine scores before eating and the eight atropine scores after eating. These comparisons resulted in a "significant" decrease of the placebo scores for both dependent variables and in a non-significant result for the atropine scores. One conclusion was that atropine prevents a lowering of the defecation threshold after eating.

Give reasons why no credence can be given to the results of this study.

6.9. Eight healthy volunteers (four females and four males, hospital staff) participated in a study (cf. [8]) on the effect of nutrient ingestion on the rectal sensitivity. Two measurements were performed on each of two days with each subject, where the two days were separated by a further day where each day was preceded by an overnight fast. On one of the two days the subjects received no meal between the two measurements, on the corresponding other day they received 600 ml of a calorically dense liquid meal between the two measurements. The time between the two measurements was set to 10 min. The order "meal on day one, no meal on day two" was chosen for two males and two females, whereas this order was inverted for the remaining four subjects. Four statistical comparisons were performed for each of the two dependent variables: 1. Comparison of the eight pretest and posttest scores in the condition without meal. 2. Comparison of the eight pretest and posttest scores in the condition with meal. 3. Comparison of the eight pretest scores for the two conditions. 4. Comparison of the eight posttest scores for the two conditions. In order to avoid the effect of an accumulation of tests an alpha-adjustment using $\alpha/4$ instead of the original α was performed. A "significant" result was found only for the second comparison for both dependent variables which led to the interpretation that nutrient ingestion increases rectal sensitivity.

Give reasons why this interpretation of the outcomes is not valid.

6.10. In a study (cf. [6]) dealing with specific psychophysiological reactions of migraine patients to stress, recovery, and relaxation, 37 migraine patients were recruited via a newspaper announcement. As a control group 44 headache-free subjects, most of which were paid, were recruited on a university campus. Note that the migraine patients were considerably older (in the mean 6.7 years!) than the control subjects.

Give reasons why it is not possible to draw causal conclusions about psychophysiological reactions of migraine patients when such a study design is being used.

6.11. In a study (cf. [7]) treating physiological response specifity and cognitive coping in migraine patients under stress, 33 patients were recruited by newspaper information. 10 control subjects were recruited from amongst the friends and relatives of the migraine patients, while 22 further paid control subjects were recruited at the university. Here, 13 control subjects had a low and 19 control subjects had a high educational standard, while 19 patients had a low and 14 patients had a high educational standard.

Give reasons why it is impossible to detect specific reactions of migraine patients with such a study design.

7 Designs without Repeated Measures

As already discussed (cf. Section 4.1.4, 4.8, 4.12, and 6.3) one had better dispense with repeated measures, i.e. with the repeated recording of scores of a dependent variable at the same subject, whenever this is possible, regarding the problem to be investigated. In particular, one should not try to "economize" on subjects by using repeated measures. In general, with an appropriate randomization, causal conclusions can be drawn, if designs without repeated measures are being used. In this case any doubt about these conclusions due to more or less obvious alternative explanations is unnecessary.

7.1 Designs with One Independent Variable

As already discussed in Section 6.5, the interpretation of outcomes which arise in experimental designs with two or more factors is being complicated due to possible interactions, i.e. the occurrence of different differential effects. In addition, larger sample sizes are required and difficulties can arise in the simultaneous realization of several factors. If one is not interested in the detection of differential effects, one had better use several one-factor designs, i.e. designs with a single independent variable instead of one multifactorial design.

C	T

Figure 7.1: One-factor design without repeated measures with one control condition (C) and one treatment condition (T)

The simplest design of this kind is depicted in Figure 7.1. Here, a sample of subjects is randomly divided into a control group (C) and a treatment group (T). If an essential difference between the scores of the dependent variable for the two groups is found, one concludes that the treatment has had an effect, which would not have occurred without the treatment. Such a conclusion is not possible if no control group is being used, as this would make any decision on whether and how the scores of the dependent variable have been influenced by the treatment impossible. In particular, a control group permits to decide whether a treatment has had a positive or negative effect.

C1	C2	...	C*k*	T

Figure 7.2: One-factor design without repeated measures with *k* control conditions (C1, ..., C*k*) and one treatment condition (T)

The interpretability of the outcomes can often be improved if more than one control condition is provided for (cf. Figure 7.2). If, e.g., the treatment is a surgical treatment, a possible control group might consist of patients, which are waiting for the treatment (**waiting control group**). For such waiting control groups there exists a risk that some patients seek medical assistance outside the study, though the organizer of the study might even not be informed. Thus, a selection might occur in the control group where those patients are no longer available as control subjects, which are either exceptionally ill or simply dissatisfied. As the patients from this particular subpopulation still form part of the treatment group, a comparison of the two groups does not permit a valid causal conclusion with respect to the effect of the treatment. It is, therefore, advisable to use a control group with a sham treatment in addition to the waiting control group. In many cases, this will be operationalized by the application of an ineffective drug, i.e. a **placebo** (cf. Section 4.10.2). As different placebos, e.g. water, saline or glucose, might have different effects, one should use different control conditions again. In animal studies **sham operations** (sham lesions) are often used as a control condition. Depending on the actual treatment it may be advantageous to introduce socially accepted sham treatments, which might at least make the patient believe that his state of health is being improved. Examples might be the use of autogenic training, yoga or homeopathy. A further step into the direction of less effective control treatments is the use of a well-known and conventional standard treatment. Here, the ethical problem arising from a non-treatment of patients, even for a restricted time, is rather small. This is in particular true, as known risks of a standard treatment, as opposed to possible unknown risks of a new treatment, are being dealt with.

T1	T2

Figure 7.3: One-factor design without repeated measures with two treatment conditions (T1 and T2)

If, as in Figure 7.3, no control is used but only two treatment conditions T1 and T2 are being compared, it is possible to draw the causal conclusion that one treatment is better than the other one, if an essential difference between the outcomes of the two groups is observed. Such a design might result if a standard treatment is used in the control condition.

C	T1	T2

Figure 7.4: One-factor design without repeated measures with two treatment conditions (T1 and T2) and one control condition (non-treatment C)

Though, in general, a causal conclusion is possible with regard to the outcomes of the design in Figure 7.3, this design should be avoided since only conclusions with respect to the relative effects but not with respect to the absolute effects of both treatments are possible. This becomes clear if one considers a control condition corresponding to a non-treatment, which is added to the design in Figure 7.3. If one assumes that treatment T2 turns out to be better than treatment T1, three different cases are conceivable:

Case 1: Both treatments are worse than the non-treatment.

Case 2: Both treatments are better than the non-treatment.

Case 3: Treatment T1 is worse and treatment T2 is better than the non-treatment.

In addition to case 1 and 2 the following case is also possible, if both treatments do not differ very much with regard to their effectiveness:

Case 4: Both treatments do not differ much from the non-treatment with respect to effectiveness.

Obviously, it is important to know, which of the four above cases is to be assumed if an appropriate comparison of the two treatments T1 and T2 is intended. Therefore, a control condition as in Figure 7.4 is indispensable.

C1	C2	...	C*k*	T1	T2

Figure 7.5: One-factor design without repeated measures with two treatment conditions (T1 and T2) and *k* control conditions (C1, ..., C*k*)

Due to the difficulty to find suited control conditions or because of the need for more than one control condition caused by this difficulty, respectively, it is natural to extend the design of Figure 7.4 under consideration of Figure 7.2 to the design in Figure 7.5.

If, in general, *m* treatments rather than two are being compared, the very common design in Figure 7.6 is considered instead of the designs in Figure 7.1 and 7.4.

C	T1	T2	...	T*m*

Figure 7.6: One-factor design without repeated measures with *m* treatment conditions (T1, ..., T*m*) and one control condition (C)

This design can be extended in analogy to the design in Figure 7.5 by providing for more than one control condition. This extension is described in Figure 7.7. A design as given in Figure 7.7 requires that the original sample of subjects is randomly split up into (*k* + *m*) groups (*k* control groups, *m* treatment groups).

C1	C2	...	C*k*	T1	T2	...	T*m*

Figure 7.7: One-factor design without repeated measures with *k* control conditions (C1, ..., C*k*) and *m* treatment conditions (T1, ..., T*m*)

In a special case of the design in Figure 7.6, which is very common, the levels of the independent variable exhibit a natural order. Such an order could arise if, in addition to a control condition, new treatment components are constantly being added in a cumulative way. Very often a drug is being applied in (m + 1) doses which can be ordered according to their size. Here, the first dose (D0) corresponds to the application of a placebo (control condition C) and the other m doses (D1, ..., Dm) are ordered with respect to increasing size.

D0 = C	D1	D2	...	Dm

Figure 7.8: One-factor design without repeated measures with (m + 1) doses (D0, D1, ..., Dm) of a drug which are ordered with respect to size, where the dose D0 corresponds to the application of a placebo (control condition C)

In Section 2.2 the aspects according to which the sizes of the doses should be selected were discussed. This specification of Figure 7.6 is depicted in Figure 7.8. The outcomes of such a design, where the original sample of subjects is randomly split up into (m + 1) subsamples permit conclusions with respect to the relation between the doses and the corresponding effects of the drug. It is thus possible to derive **dose-response curves**. These curves do not have to be monotonically increasing or decreasing, but can have, e.g., maxima or minima for intermediate doses. D0, D1, ..., Dm are not necessarily the doses of a drug. These may well be different noise levels, whose effects on the performance in a concentration task are studied.

7.2 Designs with Two Independent Variables

If the effects of two independent variables are to be studied and the mutual influence of these variables is of no interest, a one-factor design as proposed in Section 7.1 should be used for each of the two variables. This has already been discussed in Section 6.5. However, if one is interested in possibly different **differential effects**, i.e., if one wants to know whether different effects result for one of the independent variables depending on the level of the other variable, a two-factor design has to be used. The mutual influence of independent variables with respect to their effects on one or more dependent variables is called an **interaction**.

	C2	T2
C1	C1C2	C1T2
T1	T1C2	T1T2

Figure 7.9: Two-factor design with the two treatments T1 and T2 and the two control conditions C1 and C2

The simplest two-factor design arises if two treatments (T1 and T2) are being combined with the corresponding two control conditions (C1 and C2) as depicted in Figure 7.9. Here, the original sample is randomly split up into four subsamples, which are randomly assigned to the four treatment combinations C1C2, C1T2, T1C2, and

T1T2. By comparing C1C2 with T1C2 an effect caused by the first factor can be revealed, if at the same time the control condition of the second factor is present. The comparison of C1T2 with T1T2 shows, whether the first factor has had an effect if at the same time the treatment condition of the second factor is present. If the differences between T1 and C1 under C2 or T2, respectively, differ considerably, an interaction might be present. Similarly, we can compare C1C2 with C1T2 or T1C2 with T1T2, respectively, and find out, whether T2 and C2 cause different differences with respect to the dependent variable for a fixed condition C1 or T1, respectively.

C1	T1

C2	T2

Figure 7.10: Two one-factor designs with the treatments T1 and T2 and the corresponding control conditions C1 and C2

If only the effect of the first or second factor, respectively, is of interest but not possible differential effects, it is better to use two one-factor designs as depicted in Figure 7.10. In this case again the original sample would be randomly split up into four subsamples.

	B1	B2
A1	A1B1	A1B2
A2	A2B1	A2B2

Figure 7.11: Two-factor design with the treatment factor A (levels A1 and A2) and the treatment factor B (levels B1 and B2)

In analogy to Figure 7.3 we can also consider the two-factor design in Figure 7.11 where, again, the original sample has to be randomly split up into four subsamples. Here, an essential difference of the dependent variable for the combinations A1B1 and A2B1 would mean that both treatments A1 and A2 of the first factor are different if at the same time the treatment B1 of the second factor is effective. In an analogous way the outcomes for the other three comparisons A1B2 with A2B2, A1B1 with A1B2, and A2B1 with A2B2 have to be interpreted.

	B0	B1	B2
A0	A0B0	A0B1	A0B2
A1	A1B0	A1B1	A1B2
A2	A2B0	A2B1	A2B2

Figure 7.12: Two-factor design with the treatment factor A (treatments A1, A2, and control A0) and the treatment factor B (treatments B1, B2, and control B0)

The design in Figure 7.11 has the disadvantage that only statements about relative but not about absolute treatment effects are possible because no control conditions are being provided for. A direct generalization of the design in Figure 7.4 yields the more expensive design in Figure 7.12 with nine combinations of conditions, where A0 and B0 denote the respective control conditions. Here, the original sample has to be split up into nine subsamples. E.g., with the three comparisons A0B0 with A1B0, A0B0 with A2B0, and A1B0 with A2B0 one might find out whether the treatment A1 is more effective than A2 for fixed control condition B0 of the second factor, and which of the two treatments A1 and A2 is more or less effective than the control condition A0.

	B1	B2
A1	A1B1	A1B2
A2	A2B1	A2B2

C

Figure 7.13: Two-factor design with the treatment factor A (treatments A1 and A2), the treatment factor B (treatments B1 and B2), and a control condition C

One is sometimes satisfied with the design in Figure 7.13, which corresponds to the design in Figure 7.11 to which a single control condition (C) was added. This design is obtained by omitting the combinations A1B0, A2B0, A0B1, and A0B2 in Figure 7.12 and by re-designating the combination A0B0 into C. An advantage of this reduced design is that one has to randomly split up the original sample into only five instead of nine subsamples. A disadvantage is that many differential effects can no longer be studied and that, hence, no statements about the absolute effects of the treatments are possible. This can be seen in Figure 7.14, which is the original representation of Figure 7.13.

	B0	B1	B2
A0	A0B0	--	--
A1	--	A1B1	A1B2
A2	--	A2B1	A2B2

Figure 7.14: Reduction of Figure 7.12 to Figure 7.13

The design in Figure 7.13 and 7.14, respectively, is of advantage only if no essential effects can be detected in the design in Figure 7.11. In this case it is possible to test whether the four combinations have had any effect by comparing the conditions A1B1, A1B2, A2B1, and A2B2 with the control condition C = A0B0. If no essential differences between all five groups can be found, however, this does on no account mean that none of the conditions A1, A2, B1, and B2 was effective. Theoretically, it is possible that the simultaneous effect of a treatment level of a factor A (A1 or A2) and of a treatment level of a factor B (B1 or B2) can yield a mutual extinction of the respective effects. In the design of Figure 7.12 this would not be a problem since it is possible to obtain information about the effects of A1 and A2 by comparing the groups A0B0, A1B0, and A2B0. Information about the effects of B1 and B2 is obtained by a comparison of the groups A0B0, A0B1, and A0B2. The information can be obtained in a more economic way, by using six instead of nine groups and considering two one-factor designs as in Figure 7.15, thus lacking the possibility to detect different differential effects, however.

A0	A1	A2

B0	B1	B2

Figure 7.15: Two one-factor designs for the factor A (treatments A1 and A2, control A0) and the factor B (treatments B1 and B2, control B0)

	B01	...	B0c_2	B1	...	Bt_2
A01	A01B01	...	A01B0c_2	A01B1	...	A01Bt_2
⋮	⋮	...	⋮	⋮	...	⋮
A0c_1	A0c_1B01	...	A0c_1B0c_2	A0c_1B1	...	A0c_1Bt_2
A1	A1B01	...	A1B0c_2	A1B1	...	A1Bt_2
⋮	⋮	...	⋮	⋮	...	⋮
At_1	At_1B01	...	At_1B0c_2	At_1B1	...	At_1Bt_2

Figure 7.16: Two-factor design with c_1 control conditions and t_1 treatment conditions of factor A and c_2 control conditions and t_2 treatment conditions of factor B

If the t_1 treatments A1, ..., At_1 and the c_1 control conditions A01, ..., A0c_1 are present for the factor A, as well as the t_2 treatments B1, ..., Bt_2 and the c_2 control conditions B01, ..., B0c_2 for the factor B, Figure 7.12 has to be extended to Figure 7.16 with a whole of $(t_1 + c_1)\ (t_2 + c_2)$ level combinations, i.e. with the same number of samples. In all, $(t_1 + c_1)\ (t_2 + c_2)\ (t_1 + c_1 + t_2 + c_2 - 2) / 2$ differential comparisons have to be performed in this design.

7.3 Designs with more than Two Independent Variables

As already discussed in Section 6.5, not only the number of required samples increases with each additional factor, which is effective at the same time as other factors, but also the number of differential comparisons, which have to be performed. In fact, for one additional factor with k levels k times as many samples are needed as without this factor and more than k times as many differential comparisons have to be performed.

In theory, much more information about the joint effects of the different factors on the dependent variable is obtained by a multifactorial comparison yielding a better insight into the complicated causal relations than one-factor or two-factor designs. In practice, however, it will be difficult, in general, to integrate this information into easy to survey causal relationships.

C3

	C2	T2
C1	C1C2C3	C1T2C3
T1	T1C2C3	T1T2C3

T3

	C2	T2
C1	C1C2T3	C1T2T3
T1	T1C2T3	T1T2T3

Figure 7.17: Three-factor design with the three treatments T1, T2, and T3 and the corresponding control conditions C1, C2, and C3

For illustration, only the simplest case of a design with more than two independent variables is being considered here. The effect of three factors which affect a dependent variable at the same time is to be determined by this design. Each of the three factors has the two levels treatment and control. One way to depict the eight combinations of conditions for this design is depicted in Figure 7.17. The original sample of subjects is to be randomly split up into eight subsamples which are randomly assigned to the eight combinations. For the sake of simplicity, we assume that the eight sample sizes are equal. If one was only interested in finding out, whether the three treatments T1, T2, and T3 have any effects, it would be better to use the three one-factor designs with altogether six groups, which are depicted in Figure 7.18.

C1	T1

C2	T2

C3	T3

Figure 7.18: Three one-factor designs for testing which of the three treatments T1, T2, and T3 have an effect which deviates from the effect of the corresponding control conditions C1, C2, or C3

In Figure 7.19 all 28 possible comparisons for pairs of combinations as they result from Figure 7.17 are listed. In particular, by the comparisons No. 2, 9, 24, and 27 the effect of treatment T1 on the dependent variable is considered if the effect of the two other treatments is kept constant. In analogy, the effect of treatment T2 can be judged by considering the comparisons No. 1, 14, 23, and 28, and the effect of treatment T3 by considering the comparisons No. 4, 11, 17, and 22. In particular, the comparison No. 2 corresponds to the first design in Figure 7.18, the comparison No. 1 to the second design, and the comparison No. 4 to the third design.

A comparison of the effects of treatment T1 and T2 is given by the comparison No. 8. In analogy, the effects of the treatments T1 and T3 or T2 and T3, respectively, are compared by the comparisons No. 15 or 10, respectively. By the comparison No. 3 the combined effect of the treatments T1 and T2 can be judged. In analogy, by the comparison No. 6 the combination of T1 and T3 is judged, by the comparison No. 5 the combination of T2 and T3, and by the comparison No. 7 the combination of all three treatments. Whether the combined effect of T1 and T2 differs from the effect of T3 alone shows the comparison No. 19. In analogy, the results of the comparisons No. 12 and 16 are interpreted. By the comparison No. 18 it is judged, whether the combined effect of the treatments T2 and T3 is influenced by the simultaneous effect of T1. In analogy, the results of the comparisons No. 13 and 25 are interpreted.

Finally, by the comparisons No. 20, 21, and 26 it is possible to decide whether the difference of two treatments is changed by the simultaneous effect of the respective third treatment.

No.	Comparison	No.	Comparison	No.	Comparison
1	C1C2C3 variables.	11	C1T2C3 vs. C1T2T3	21	T1T2C3 vs. T1C2T3
2	C1T2C3	12	C1T2C3 vs.T1C2T3	22	T1T2C3 vs. T1T2T3
3	C1C2C3 vs. T1C2C3	13	C1T2C3 vs. T1T2T3	23	C1C2T3 vs. C1T2T3
4	C1C2C3 vs. T1T2C3	14	T1C2C3 vs. T1T2C3	24	C1C2T3 vs. T1C2T3
5	C1C2C3 vs. C1C2T3	15	T1C2C3 vs. C1C2T3	25	C1C2T3 vs. T1T2T3
6	C1C2C3 vs. C1T2T3	16	T1C2C3 vs. C1T2T3	26	C1T2T3 vs. T1C2T3
7	C1C2C3 vs. T1C2T3	17	T1C2C3 vs. T1C2T3	27	C1T2T3 vs. T1T2T3
8	C1C2C3 vs. T1T2T3	18	T1C2C3 vs. T1T2T3	28	T1C2T3 vs. T1T2T3
9	C1T2C3 vs. T1C2C3	19	T1T2C3 vs. C1C2T3		
10	C1T2C3 vs. T1T2C3	20	T1T2C3 vs. C1T2T3		
	C1T2C3 vs. C1C2T3				

Figure 7.19: All 28 possible pairwise comparisons of two out of the eight possible combinations in Figure 7.17

Many other ways to interpret the outcomes are obtained, if pairs of comparisons are considered in Figure 7.19 and the differences for different comparisons are compared with each other. Altogether, we may consider here 378 different pairs of pairs. Here, we consider only one example of such a comparison: in case of an essentially larger difference of the effects on the dependent variable for the comparisons No. 9 and 24 it can be concluded that the effect of treatment T1 is not the same, if at the same time T2 or T3 is effective.

If one tries to perform in this way all possible comparisons of the outcomes of comparisons it is advisable to restrict oneself to an easy to survey subset of comparisons because of the large number of possible statements. This subset of comparisons should be chosen on the basis of hypotheses which were formulated in advance. If for certain comparisons outcomes are obtained for which different interpretations are possible, the interpretability can be improved by including further comparisons. A statistical evaluation of the results will be possible only for very few comparisons which had been fixed in advance according to founded hypotheses because otherwise we would have the problem of too many significance tests (problem of multiple testing, cf. Section 3.1.3). All other comparisons are only performed for facilitating in a descriptive way the interpretation of the results of the significance tests.

Our argumentation above may not be well understood by those researchers who are accustomed to evaluate multifactorial designs by means of analyses of variance. With respect to the example above such researchers would argue that a three-factor analysis of variance with respect to the three factors F_1 (= treatment T1, control C1), F_2 (= treatment T2, control C2), and F_3 (= treatment T3, control C3) would show for which factors (F_1, F_2 or F_3) there are **main effects** (F_1, F_2 or F_3), first-order interactions ($F_1 \times F_2$, $F_1 \times F_3$ or $F_2 \times F_3$) or a second-order interaction ($F_1 \times F_2 \times F_3$). Altogether, the results of seven significance tests had to be interpreted. Again the problem of multiple testing arises. In addition, this analysis yields only valid results, if the assumptions of an analysis of variance (independence of all scores, normally distributed scores with the same population means within each combination of conditions, equal variances for all combinations of conditions) are fulfilled.

A major shortcoming of an evaluation by means of an analysis of variance is that in contrast to the proceeding described above, the outcomes for single combinations of conditions are not compared directly but that before any comparison first the scores corresponding to several combinations are pooled. This has the consequence that significant main effects, i.e. the effects of single factors, no longer have an easy interpretation if the corresponding factors interact with other factors. Further, significant interactions do not permit an interpretation in substance without a preceding detailed analysis as it was considered by us above, because the existence of interactions only indicates that different treatments influence each other mutually but the exact nature of this influence is not known. This is true also if in addition to an analysis of variance multiple comparisons are performed which is possible for factors with more than two levels. These comparisons which can be performed for each main effect and for each interaction again are based on means of results for combinations of conditions, i.e. their interpretation may be problematic.

It is difficult to give general advice for the interpretation of the results of analyses of variance because, in principle, each possible pattern of main effects and interactions may occur. In the example above with three possible main effects and four possible interactions, altogether 128 different patterns of effects are possible, from the case "no significant effect" to the case "seven significant effects".

In order to illustrate differences of the two ways of interpreting the outcomes of multifactorial designs we fix in an arbitrary way fictitious population means for several situations. We have already used this kind of reasoning in Figure 6.6 in Section 6.5. Such fictitious population means can be imagined by assuming that we sample for each combination of conditions 1.000.000 subjects and compute the mean of the corresponding scores for each combination. Even if these means are still not totally stable, we can expect that the fluctuations, which they would show in case of a repetition of the sampling in contrast to the true population means, are small.

	Fictitious Population Means									
Level Combinations	A	B	C	D	E	F	G	H	I	K
C1 C2 C3	10	15	17	17	4	11	22	4	16	12
C1 C2 T3	10	15	1	17	30	9	22	26	14	20
C1 T2 C3	10	15	29	3	12	9	8	26	14	30
C1 T2 T3	10	15	13	3	–6	11	8	4	16	–2
T1 C2 C3	10	5	7	3	–6	9	–2	–6	4	–10
T1 C2 T3	10	5	–9	3	12	11	–2	16	6	–6
T1 T2 C3	10	5	19	17	30	11	12	16	6	40
T1 T2 T3	10	5	3	17	4	9	12	–6	4	–4

Figure 7.20: Fictitious population means for the eight level combinations from Figure 7.17 for ten different situations A, B, C, D, E, F, G, H, I, and K

In Figure 7.20 fictitious population means are given for ten different situations for each of the eight combinations in Figure 7.17. Here, A corresponds to a situation without main effects and without interactions, B to a situation without interactions and only one main effect F_1 (more precise: a main effect for the factor F_1), C to a situation with three main effects F_1, F_2, and F_3, but without interactions, D to a situation without main effects, but with a first-order interaction $F_1 \times F_2$ (more precise: an interaction between the two factors F_1 and F_2), and E to a situation without main effects, without a second-order interaction, but with three first-order interactions $F_1 \times F_2$, $F_1 \times F_3$, and $F_2 \times F_3$. In situation F only a second-order interaction $F_1 \times F_2 \times F_3$ is present, but neither main effects nor first-order interactions, in situation G only a main effect F_1 and a first-order interaction $F_1 \times F_2$ is present, in situation H only a main effect F_1 and a first-order interaction $F_2 \times F_3$ is present, in situation I only a main effect F_1 and a second-order interaction $F_1 \times F_2 \times F_3$ is present, and, finally, in situation K main effects F_1, F_2, and F_3, first-order interactions $F_1 \times F_2$, $F_1 \times F_3$, and $F_2 \times F_3$, and a second-order interaction $F_1 \times F_2 \times F_3$ are present.

In Figure 7.21 the differences in the fictitious population means from Figure 7.20 for the 28 comparisons in Figure 7.19 are given. Here, for each comparison the value for the first level combination is subtracted from the value of the second level combination. For situation A only zero differences result because no effects are present. By the comparisons No. 2, 9, 24, and 27 it is tested, whether the effect of treatment T1 is different from the effect of the control condition C1, if at the same time the level combinations of the two other factors (C2C3, T2C3, C2T3, and T2T3) are kept constant. For situation B for each of these comparisons the same difference (–10) results which is smaller than zero. This means that independent of the levels of the factors F2 and F3 the treatment T1 yields always a value of the dependent variable which is smaller than the value for the control condition C1 by a value of 10. Because the mean of the four equal differences, of course, also equals (–10) and is, therefore, different from zero, a main effect of factor F_1 is present according to the definition of main effects in the analysis of variance. For the comparisons No. 1, 14, 23, and 28 we get for situation B always the same difference 0, i.e. no effect of the treatment T2 with respect to the control condition C2 does exist. Because the mean of these four zero differences is also zero, no main effect of factor F_2 is present. The same result we find in situation B for factor F_3, if the comparisons No. 4, 11, 17, and 22 are considered. It is decisive that all of the respective four comparisons yield the same difference, which has the consequence that the interpretation in substance of the main effect, e.g., for the factor F_1, coincides with the interpretation of the effect for each of the four comparisons No. 2, 9, 24, and 27 separately. With respect to the factor F_1 we find analogous results also for the situations C and H, i.e. for situations where the factor F_1 does not interact with the factors F_2 or F_3.

In situation D the four comparisons No. 2, 9, 24, and 27 yield the differences –14, 14, –14, and 14, i.e. partly the same effects of treatment T1, partly just the opposite with respect to the control condition C1 depending on the level combinations of the two other factors F_2 and F_3. Obviously, in this situation also a differential effect of treatment T1 must be assumed. However, the mean of the four differences is zero, i.e. an analysis of variance would find no effect of factor F_1. Analogous results are found in situations E and F.

In situation G the four comparisons No. 2, 9, 24, and 27 yield the differences –24, 4, –24, and 4. Depending on the level combination of the factors F_2 and F_3 we find larger values of the dependent variable either for treatment T1 or for the control condition C1. The mean of the four differences is here (-10), i.e. an analysis of variance would detect a (negative) main effect of factor F_1 though in the presence of the level combinations T2C3 or T2T3 treatment T1 yields larger values than the control condition C1 which yields a positive difference (4). Again, a misleading interpretation of the results is found by considering the main effect of the factor F_1 detected by an analysis of variance. Analogous results are obtained for the situations H, I, and K.

No.	A	B	C	D	E	F	G	H	I	K
1	0	0	12	-14	8	-2	-14	22	-2	18
2	0	-10	-10	-14	-10	-2	-24	-10	-12	-22
3	0	-10	2	0	26	0	-10	12	-10	28
4	0	0	-16	0	26	-2	0	22	-2	8
5	0	0	-4	-14	-10	0	-14	0	0	-14
6	0	-10	-26	-14	8	0	-24	12	-10	-18
7	0	-10	-14	0	0	-2	-10	-10	-12	-16
8	0	-10	-22	0	-18	0	-10	-32	-10	-40
9	0	-10	-10	14	18	2	4	-10	-8	10
10	0	0	-28	14	18	0	-14	0	0	-10
11	0	0	-16	0	-18	2	0	-22	2	-32
12	0	-10	-38	0	0	2	-10	-10	-8	-36
13	0	-10	-26	14	-8	0	4	-32	-10	-34
14	0	0	12	14	36	2	14	22	2	50
15	0	10	-6	14	36	0	24	32	10	30
16	0	10	6	0	0	2	10	10	12	8
17	0	0	-16	0	18	2	0	22	2	4
18	0	0	-4	14	10	0	14	0	0	6
19	0	10	-18	0	0	-2	10	10	8	-20
20	0	10	-6	-14	-36	0	-4	-12	10	-42
21	0	0	-28	-14	-18	0	-14	0	0	-46
22	0	0	-16	0	-26	-2	0	-22	-2	-44
23	0	0	12	-14	-36	2	-14	-22	2	-22
24	0	-10	-10	-14	-18	2	-24	-10	-8	-26
25	0	-10	2	0	-26	0	-10	-32	-10	-24
26	0	-10	-22	0	18	0	-10	12	-10	-4
27	0	-10	-10	14	10	-2	4	-10	-12	-2
28	0	0	12	14	-8	-2	14	-22	-2	2

Figure 7.21: Differences of the fictitious population means from Figure 7.20 for the 28 comparisons from Figure 7.19

Total (as in situations D, E, and F) or partial (as in situations H, I, and K) misinterpretations of main effects detected by analyses of variance can only occur if interactions are present. E.g., a first-order interaction ($F_1 \times F_2$) for the two factors F_1 and F_2 or a second-order interaction ($F_1 \times F_2 \times F_3$) for the three factors F_1, F_2, and F_3 is always present if the differences for the four comparisons No. 2, 9, 24, and 27 are not all equal. If for the comparisons No. 2, 9, 24, and 27 and for the comparisons No. 1, 14, 23, and 28 and the comparisons No. 4, 11, 17, and 22, respectively, always equal differences result, as this is the case for the situations A, B, and C, neither first- nor second-order interactions are present.

A first-order interaction $F_1 \times F_2$ between the two factors F_1 and F_2 exists, if the mean of the differences for the comparisons No. 2 and 24 is different from the mean for the comparisons No. 9 and 27, as this is the case for the situations D, E, G, and K. For example, we find in situation K for the first mean $(-22 - 26) / 2 = -24$ and for the second mean $(10 - 2) / 2 = 4$.

A second-order interaction $F_1 \times F_2 \times F_3$ between the three factors F_1, F_2 and F_3 exists, if the difference of the differences for the two comparisons No. 2 and 9 is different from the difference of the differences for the two comparisons No. 24 and 27, as this is the case for the situations F, I, and K. For example, we find in situation K for the first difference the value $(-22) - 10 = -32$, and for the second difference the value $(- 26) - (-2) = -24$.

If one compares for multifactorial designs the interpretational approach via the direct comparisons of the dependent variable for two respective level combinations with the analysis of variance approach via main effects and interactions, the elementary first approach seems to give not only more simple but also more appropriate interpretations.

Summary

1. The best interpretation is possible for outcomes from one-factor designs without repeated measures, where one or more control groups are present.

2. In designs with two or more factors the existence of interactions cannot be ruled out. If one is not interested in detecting interactions it is better to use several one-factor designs instead of one multifactorial design.

3. Also in case of multifactorial designs control groups should be present.

4. The interpretation of the outcomes of two-factor designs or multifactorial designs by means of an analysis of variance is only appropriate in special cases and it is better to perform direct comparisons of the dependent variable for pairs of level combinations.

Questions

7.1. Why is it not possible to "save" subjects by using repeated-measures designs?

7.2. What are the advantages and disadvantages of one-factor designs?

7.3. Why are control groups needed?

7.4. Give an example for a one-factor design where it would be advantageous to have four different control groups.

7.5. What are relative and absolute effects of treatments?

7.6. Formulate a concrete design corresponding to the design in Figure 7.7.

7.7. Explain by means of a concrete example what is understood by a dose-response curve and in which way it is obtained.

7.8. Give a concrete example for an interaction in a two-factor design.

7.9. Compare the three designs in the Figures 7.12, 7.13, and 7.15 with respect to their advantages and disadvantages.

7.10. Describe a concrete design according to Figure 7.16, where $c_1 = 1$, $t_1 = 2$, $c_2 = 3$, and $t_2 = 4$.

7.11. Give an example for a three-factor design, in which the first factor has two levels, the second factor three levels, and the third factor four levels.

How many samples are necessary and how many differential comparisons can be performed?

7.12. Give an example for a four-factor design, where each factor has two levels.

How many samples are necessary and how many differential comparisons can be performed?

7.13. How many pairwise comparisons of the form depicted in Figure 7.19 could be considered for the designs in Question 7.11 and 7.12?

7.14. Explain, in which way main effects, first- and second-order interactions are to be interpreted in the three-factor design in Question 7.11.

7.15. Which main effects and interactions are possible in the design considered in Question 7.12?

7.16. How many possible analysis of variance tests (without multiple comparisons) can be considered for the designs in Question 7.11 and 7.12?

7.17. How many different "effect patterns" as a result of an analysis of variance are possible for the design in Question 7.12?

8 Designs with Repeated Measures

The problems which arise if more than one measurement is performed with a subject (repeated measures) have been discussed repeatedly throughout this book (cf. Section 4.1.4, 4.8, 4.12, 6.3, and 6.4). In general, the claimed advantages of repeated-measures designs do not justify the use of this kind of designs in view of the profound disadvantages. In most cases they render a causal interpretation of the outcomes impossible—due to the many possible alternative explanations. Though rare in practice, problems can be constructed, which can only be studied by means of repeated-measures designs. One possible example is a study investigating whether the effect of a measurement has been influenced by a preceding measurement at the same subject. Designs for this kind of problems require more experimental groups and, thereby, also a larger number of subjects in comparison to designs without repeated measures if one intends to use the outcomes to establish causal relationships.

8.1 Designs with One Independent Variable

If one investigates whether the intake of an anorectic leads to a loss of weight, a first study design might be to apply the drug to a random sample of subjects and to measure the weight a week later. If the weight of the subjects is lower on the average than the normal weight in the population, the effectiveness of the drug might be assumed. However, such a study design, depicted in Figure 8.1, can, obviously, not be recommended, since one cannot rule out that the drug has had no effect at all or may even have caused an increase in weight. This might have been due to various reasons: the sample might not have been a real random sample and it might have yielded scores, which are systematically below (or above) the normal scores. On the other hand it might have been a random sample but it might, nevertheless, have yielded scores, which were below (or above) the normal scores on the average. What is more, the normal reference score could have been out of date or cannot be trusted for other reasons; or the subjects might have changed their eating behavior, as they know that they have taken a drug, etc.

T		A

Figure 8.1: **One group posttest only design:** After a treatment (T) follows a period which might be very short, without treatment or measurement, and after this a posttest (A)

B	T		A

Figure 8.2: **One group pretest-posttest design**: After a pretest (B) follows a treatment (T), then a period which might be very short, without treatment or measurement, and after this a posttest (A)

A spurious improvement of the design in Figure 8.1 is obtained if the weight of the subjects is being measured not only after the application of the drug but also directly before the application. In this case, a **repeated-measures design** is used, which, here, in particular, is a **before-after design** or a **pre-post design** (cf. Figure 8.2). If the posttest scores of the weight are lower than the pretest scores on the average, one might conclude that the drug has been effective. However, this conclusion is difficult to justify in view of many alternative explanations. If, e.g., the subjects know that they participate in a study about the effectiveness of an anorectic, they might probably control their food intake during the study more consciously than they would usually do. Weight loss might result solely due to this self-control. This effect is reinforced by the fact that, in particular, subjects, who at least on a subjective scale are discontent with their weight, are inclined to participate in such a study. However, if the subjects do not know that they get an anorectic, which might cause ethical problems, they at least know that they participate in a drug study. As a consequence they will observe their physical and psychic condition during the study with much more attention than they usually would. In particular, they will notice everything that can affect their condition, e.g. the food intake. Again, a decrease of weight may be solely caused by control processes, which are induced by participating in the study.

B_1	B_2	...	B_m	T		A

Figure 8.3: One-group **time-series design** with **baseline** and **intervention**: After m (with $m \geq 2$) pretests (B_1, ..., B_m) at m successive points of time follows a treatment (T), after this a time interval without treatment or measurement, which may be very short, and after this a posttest (A)

The same argument holds, if the before-after design is extended to a **time-series design** with **intervention** (cf. Figure 8.3) where not only one weight score but weight scores for m days (with $m \geq 2$) (the so-called **baseline**) are recorded before the application of the drug. An advantage of this procedure is that it is easier to evaluate the normal fluctuation of a subject's weight than with only one pretest. Consequently, one should not only use more than one pretest but also more than one posttest (cf. Figure 8.4).

B_1	B_2	...	B_m	T		A_1	A_2	...	A_n

Figure 8.4: One-group time-series design with **baseline**, **intervention** and post-treatment phase: After m (with $m \geq 2$) pretests (B_1, ..., B_m) at m successive points of time a treatment (T) follows, after this a time interval without treatment and measurement, which may be very short, and after this n (with $n \geq 2$) posttests (A_1, ..., A_n) at n successive points of time

One might, of course, also want to render the alternative explanation given above implausible by ensuring, that the subjects do not know that they are participating in a study. For this, the anorectic has to be inserted into their daily meals and the weight could be registered any time the subjects step on a certain plate in the floor. Apart from the ethical and practical problems with such a procedure, only one of many alternative explanations could be made implausible by this. A smaller food intake and thus a decrease in weight due to higher outdoor temperatures or an altered diet would, by no means, be ruled out as an alternative explanation by the described procedure.

Group 1	T		A
Group 2	C		A

Figure 8.5: Two-group design in analogy to the one-group design in Figure 8.1, where C corresponds to a control condition (e.g., a placebo)

Group 1	B	T		A
Group 2	B	C		A

Figure 8.6: Two-group design in analogy to the one-group design in Figure 8.2, where C corresponds to a control condition (e.g., a placebo)

Group 1	B_1	B_2	...	B_m	T		A
Group 2	B_1	B_2	...	B_m	C		A

Figure 8.7: Two-group design in analogy to the one-group design in Figure 8.3, where C corresponds to a control condition (e.g., a placebo)

Group 1	B_1	B_2	...	B_m	T		A_1	A_2	...	A_n
Group 2	B_1	B_2	...	B_m	C		A_1	A_2	...	A_n

Figure 8.8: Two-group design in analogy to the one-group design in Figure 8.4, where C corresponds to a control condition (e.g., a placebo)

A causal conclusion can only be drawn if a control condition is being introduced. In the example with the anorectic this might be a placebo, i.e. an ineffective pseudo-drug (cf. Section 4.10.2). For this, the original sample is being randomly split up into two subsamples, one of which receives the anorectic and the other the placebo. If the posttest scores of the two groups differ considerably, an effect of the anorectic can be assumed. The corresponding two-group designs are depicted in Figure 8.5, 8.6, 8.7, and 8.8. If more than one posttest score is being recorded, as in Figure 8.8, an average of the posttest scores can be computed for each subject, in order to obtain a more stable effect measure. In this case, the samples of means of the posttest scores are compared with each other.

In Figure 8.7 and 8.8 the number (m) and the time pattern of the pretest scores (B_1, ..., B_m) and also the number (n) and the time pattern of the posttest scores (A_1, ..., A_n)

were chosen to be the same for both groups, as one otherwise could not rule out that differences between the posttest scores of the two groups are found which do not result from a different effect of treatment (T) and control condition (C).

It is obvious that the pretest scores in the two-group designs in Figure 8.5, 8.6, 8.7, and 8.8 are not vital to draw causal conclusions about the effectiveness of the drug. Therefore, these measurements are superfluous. Rather, one cannot rule out that a measurement of weight before the application of the anorectic or the placebo, respectively, can even have the effect that subjects are sensitized (cf. Section 3.2.3), i.e. influenced in their eating behavior. A possible effect of such a sensitization might be that the posttest scores of the two groups do not differ very much. Similarly, the common belief that pretests can be used in order to increase the evidence of posttest scores must be rejected on the basis of many arguments (cf. Section 4.13).

B	T	A	B	T	A	B	T	A	B	T	A

Figure 8.9: One-group design, where each subject is exposed to the sequence pretest (B), treatment (T), posttest (A) four times

One might want to believe that a more complicated design with repeated measures would permit more conclusive interpretations. E.g., assume that for a group of subjects the weight is recorded daily for 12 weeks. In the second, fifth, eighth, and eleventh week of the study the subjects get the anorectic daily, and do not get any drugs during the other weeks. The scores of the first week are being used as a baseline and the scores of the third week as posttest scores for the treatment in the second week. Similarly, we use the scores of the fourth week as pretest scores and the scores of the sixth week as posttest scores for the treatment in the fifth week. In the eighth and eleventh week we proceed analogously with the treatment (cf. Figure 8.9). In this design the posttest scores of a treatment are not considered as the pretest scores for the subsequent treatment, since otherwise the effects of the treatment at different points of time might not be distinguishable. As a consequence of this, pretest and posttest scores might not differ in spite of a treatment effect.

B	T	A		B	T	A		B	T	A		B	T	A

Figure 8.10: One-group design, where each subject is exposed to the sequence pretest (B), treatment (T), posttest (A) four times, and where the four sequences are separated by three periods without treatment or measurement

It might be sensible to add a **wash-out period** without treatment or measurement between each posttest (A) and each following pretest (B) in Figure 8.9, as depicted in Figure 8.10. This helps to avoid a superposition of the effects of the treatment at different points of time.

For the designs in Figure 8.9 or 8.10 one might argue that the anorectic has lead to the desired effect if for each of the four treatment periods the posttest results are below those of the pretests. But a possible alternative explanation might be that the subjects have been sensitized with respect to their weight each time they received the drug in connection with the daily weighing. Thus, the treatment could have caused a change in behavior, which in turn produced the observed effect though the drug itself had no physiological effect.

B	T	A	B	C	A	B	T	A	B	C	A	B	T	A	B	C	A	B	T	A	B	C	A

Figure 8.11: One-group design where treatment (T) and control condition (C) are used four times together with pretests (B), posttests (A), and periods without treatments or measurements placed between posttests and pretests

One could try to render this possible alternative explanation implausible by introducing a control condition into the design in Figure 8.10, as depicted in Figure 8.11. This design would comprise 31 weeks if we fix the duration of the seven periods without treatment and measurement to one week each. If one assumes that treatment and control condition cannot be discerned by the subjects, which might be achieved by choosing a placebo which does not differ from the anorectic with respect to appearance, taste, and odor, and if the posttest scores after the control condition are always higher than the posttest scores after the treatment this is a strong indication of the effectiveness of the anorectic. As a rule, one must expect the outcomes to be not as distinct as one would want them to be, as either the treatment effect decreases with time (decreasing trend) or the effects of succeeding treatments accumulate (increasing trend). This can have the consequence that differences between treatment and control condition are not constant over time. Further, it is difficult to refute the argument that subjects might perceive the difference between drug and placebo in some way—independent of the mere physiological effect.

If the study is not performed as a double-blind study, an experimenter effect (cf. Sections 3.3.6, 3.3.7, and 4.10.5) cannot be ruled out. It is also possible that the strict alternation of drug and placebo is perceived, in particular, if the subjects have been informed about the aim of the study in advance. Finally, one cannot rule out that the chosen alternating sequence of utilizing drug and placebo coincides with possible natural feeding rhythms of the subjects, caused, e.g., by the specific menu in a canteen.

All these alternative explanations could be ruled out by the far more simple two-group design in Figure 8.5, which, therefore, is to be preferred in all situations. Here it becomes obvious that the argument that it is possible to "save" subjects using the design in Figure 8.11 instead of the design in Figure 8.5 is not realistic. First, one "loses" all participating subjects in a study, which does not permit a causal conclusion, which is not acceptable for ethical and financial reasons. Second, one has to expect for a design with a long duration as depicted in Figure 8.11, that many subjects will abandon the study before its official end. As the outcomes of such **dropouts** due to **experimental mortality** (cf. Section 3.2.7) usually cannot be considered in an appropriate way in the evaluation of the study, either the outcomes of additional subjects must be recorded subsequently or more subjects must be included from the start. Both procedures can result in a far higher need of subjects as in the simpler design.

A further argument against the design in Figure 8.11 concerns external validity (cf. Chapter 3). First, subjects are far more willing to participate in a study with a short duration with a design that is depicted in Figure 8.5, than in a study with a long duration as in Figure 8.11. Thus the subjects participating in long-term studies and in short-term studies, respectively, form two different groups. Due to this kind of selection, any conclusion from the study of long duration can, therefore, only be generalized to a usually very small population of subjects, which have the time to participate in a study which lasts many weeks. Second, an additional restriction of generalizability occurs if dropouts are present because it can be assumed that subjects, which participate in a study of long duration up to the end differ from subjects who leave the study before its end. All in all, this example demonstrates that in repeated-measures designs not only threats to internal validity may be present, i.e. it is impossible to draw causal conclusions, but also threats to external validity may occur, i.e. it is difficult to generalize any effects to a larger population.

In view of the interpretational problems with repeated-measures designs with only one treatment, it is obvious that these problems are aggravated in the case of two or three treatments. With respect to our example, assume that the effects of two anorectics (T1 and T2) are to be compared. If only the comparison of T1 and T2 is of interest, the designs in Figure 8.5, 8.6, 8.7, 8.8, and 8.11 can be directly adopted by identifying T1 with T and T2 with C. One had better use a control condition C in addition to the treatments T1 and T2. Thus, there will not be any interpretational problems if a sample of subjects is randomly split up into three groups and the three-group design of Figure 8.12 is being used.

Group 1	T1		A
Group 2	T2		A
Group 3	C		A

Figure 8.12: Three-group design with two treatments (T1 and T2), a control condition (C), and possibly very short time periods without a treatment condition (T1, T2 or C) or a measurement and which are succeeded by a posttest (A)

In repeated-measures designs all experimental conditions (T1, T2, and C) may be presented to a group of subjects in a suited sequence—see the example depicted in Figure 8.13. In this design each of the three conditions (T1, T2 or C) has each of the corresponding two other conditions once as a predecessor and once as a successor. Here, it is problematic, e.g., that after-effects which affect not only the immediately following condition may yield outcomes which are difficult to interpret. Other problems have already been discussed in the context of the designs in Figure 8.9, 8.10, and 8.11.

T1		A	T2		A	C		A	T2		A	T1		A	C		A	T1		A

Figure 8.13: Repeated-measures design with two treatments (T1 and T2), a control condition (C), and possibly very short time periods without a treatment condition (T1, T2 or C) or a measurement and which are succeeded by a posttest (A)

Another approach considers independent groups for different sequences of the experimental conditions as demonstrated in the six-group-crossover design in Figure 8.14. The design in Figure 8.14 is not only much more large-scale than the design in Figure 8.12, but, moreover, the outcomes of this design cannot easily be interpreted if we do not restrict ourselves to considering only the first experimental condition of the sequence for each group, i.e. the first posttest score (A). Then, the outcomes for groups 1 and 2, 3 and 4, and 5 and 6 can be pooled, such that a design results which is equivalent to the one in Figure 8.12.

Group 1	C		A	T1		A	T2		A
Group 2	C		A	T2		A	T1		A
Group 3	T1		A	C		A	T2		A
Group 4	T1		A	T2		A	C		A
Group 5	T2		A	C		A	T1		A
Group 6	T2		A	T1		A	C		A

Figure 8.14: Six-group-crossover design with two treatments (T1 and T2), a control condition (C), and possibly very short time periods without a treatment condition (T1, T2 or C) or a measurement and which are succeeded by a posttest (A)

Some problems of the interpretation of the outcomes of such crossover designs have already been discussed in Section 4.8 and 6.4. If one, e.g., assumes that the treatments T1 and T2 have irreversible effects where a single application is enough for the organism to no longer react to the corresponding other treatment, no conclusion with respect to the treatments can be drawn from the third posttest in group 1 and 2, and from the second and third posttest in group 3, 4, 5, and 6.

We have to point out that neither crossover designs nor other repeated-measures designs are necessary if one is interested in the effects which result if one or more treatments are being applied in a sequence, i.e. if the after-effects of a treatment on following treatments are of interest. The simplest questions, which may be asked here, are the following: Do the effects cumulate? Is there first an accumulation followed by a saturation? Is there a fading of the effects? Are there irreversible effects?

In Figure 8.15 a design is depicted which one can use to investigate the additional effect of a further treatment if the same treatment is utilized at each point of time. A comparison of the posttests of group 1 and 2 shows whether there is an effect at all, while a comparison of the posttests of group 3 and 4 shows which effect results from a second treatment following a first treatment. Similarly, a comparison of the posttests of group 5 and 6 reveals the effect of a third treatment after two preceding treatments, while the comparison of the posttests of group 7 and 8 informs about the effect of the fourth treatment after three treatments.

In Figure 8.16 a seven-group design is depicted by which one can investigate the effect of a treatment T1 on a subsequent treatment T2 and vice versa. The comparison of group 4 with group 5 reveals the additional effect caused by treatment T2 following treatment T1. The comparison of group 5 with group 1 shows the after-effect caused by treatment T1 if no treatment follows. The direct effect of treatment T1 is revealed if group 1 is compared to group 3. With respect to the after-effects of treatment T2 on a subsequent treatment T1 these can be similarly evaluated by a comparison of group 6 and group 7, group 7 and group 2, and group 2 and group 3.

Group 1	T		A						
Group 2	C		A						
Group 3	T		T		A				
Group 4	T		C		A				
Group 5	T		T		T		A		
Group 6	T		T		C		A		
Group 7	T		T		T		T		A
Group 8	T		T		T		C		A

Figure 8.15: Eight-group design to study the effects of successive treatments (T) with a control condition (C), posttests (A), and possibly short periods without treatment and without measurement

Group 1	T1		A		
Group 2	T2		A		
Group 3	C		A		
Group 4	T1		T2		A
Group 5	T1		C		A
Group 6	T2		T1		A
Group 7	T2		C		A

Figure 8.16: Seven-group design to study the effect of a successive treatment after a preceding treatment (T1 or T2, respectively) with a control condition (C), posttests (A), and possibly short periods without treatment and without measurement

Up to now the disadvantages which arise from repeated-measures designs have been discussed. Furthermore suggestions have been made how these designs can be replaced by designs without repeated measures which permit a causal interpretation. However, if actually the influence of a measurement on subsequent measurements is to be studied, it is obvious that repeated-measures designs are needed. For the special case that the influence of a pretest on a posttest after a treatment is of interest we have already discussed the four-group design of Solomon in Section 4.10.6.

The simplest situation in this context is the case of one dependent variable measured at two different points of time (t_1 and t_2 with $t_1 < t_2$) with measurements M_1 and M_2, without any intervention by the experimenter. A design which permits to conclude whether measurement M_1 influences measurement M_2 is depicted in Figure 8.17. The comparison of M_2 with $M_2^{(1)}$ shows, whether the first measurement has had an influence on the second measurement or not. The comparison of M_1 with $M_2^{(1)}$ shows, whether only measurement M_1 influences measurement M_2 or whether the latter is also or possibly solely influenced by interim events as maturation or history (cf. Section 3.2.1 and 3.2.2).

	Point of Time	
Group	t_1	t_2
1	M_1	M_2
2		$M_2^{(1)}$

Figure 8.17: Two-group design for studying the influence of a measurement (M_1) on a succeeding measurement (M_2) without an intervention by the experimenter

If measurements are recorded at three points of time (t_1, t_2, and t_3 with $t_1 < t_2 < t_3$), the four-group design in Figure 8.18 can be used. The comparison of $M_2^{(1)}$ with M_2 indicates whether M_1 influences M_2, while the comparison of $M_3^{(2)}$ with $M_3^{(1)}$ indicates, whether M_2 has an influence on M_3. If both effects differ, one can conclude that the influence of a measurement on the subsequent measurement can be altered by interim events.

The comparison of $M_3^{(1)}$ with M_3 indicates whether the measurement M_1 also influences the measurement M_3. Such an effect may be due to a direct after-effect of M_1. But it is also possible that such an effect occurs though M_1 only affects M_2 but that the alterated M_2 has an effect on M_3. A comparison of $M_3^{(3)}$ with $M_1^{(3)}$ indicates whether M_1 has a direct influence on M_3, while a comparison of the comparisons $M_3^{(2)}$ with M_3, of $M_3^{(2)}$ with $M_3^{(1)}$, and of $M_3^{(3)}$ with $M_1^{(3)}$ indicates if the combined effect of M_1 and M_2 results from an addition of the single effects.

More complicated situations result if interventions of an experimenter occur, e.g. treatments between the measurements, since in this case the effects might have been caused not only by pretests but also by treatments. An example is the **Four-group design** by **Solomon** in Section 4.10.6 reproduced in Figure 8.19. This presentation differs from the presentation in Figure 4.22, among other things, in the fact that simply the lapse of time is taken as the control condition. A comparison of $M_2^{(3)}$ with $M_2^{(2)}$ or of $M_2^{(1)}$ with M_2, respectively, indicates if there has been a treatment effect. If the effects in both comparisons differ, an additional effect of the pretest M_1 is present. This effect can be estimated by comparing $M_2^{(3)}$ with $M_2^{(1)}$ or $M_2^{(2)}$ with M_2, respectively. If the effects of the pretest M_1 are different for the two comparisons, this indicates that the presence or absence, respectively, of the treatment alters the effect of the pretest.

	Point of Time		
Group	t_1	t_2	t_3
1	M_1	M_2	M_3
2		$M_2^{(1)}$	$M_3^{(1)}$
3			$M_3^{(2)}$
4	$M_1^{(3)}$		$M_3^{(3)}$

Figure 8.18: Four-group design used in order to study the influence of one (M_1) or two (M_1, M_2) measurements on one or two succeeding measurements without an intervention by the experimenter

	Point of Time		
Group	t_1	t_2	t_3
1	M_1	T	M_2
2	$M_1^{(1)}$		$M_2^{(1)}$
3		T	$M_2^{(2)}$
4			$M_2^{(3)}$

Figure 8.19: Four-group design of Solomon with a treatment (T), pretest (M_1), and posttest (M_2)

Probably, one is rarely interested in studying the effect of measurements on succeeding measurements only, as in the designs in Figure 8.17 and 8.18. As we have already seen, it is, in general, better to dispense with pretests, and also designs like the one in Figure 8.19 will not be of great practical relevance. The real importance of repeated-measures designs is seen in a far more common situation. For this, we consider the case where the effect of a treatment (T) is to be established by a comparison with a control condition (C). Subjects are randomly split up into two groups, where in one group T and in the other group C is being applied. After this, scores of a dependent variable are recorded and the two samples of these scores are compared.

However, in most studies this procedure is modified. Instead of the scores of only one dependent variable, in most cases the scores of several dependent variables are recorded in a given order, e.g., blood pressure, heart rate, and state of health. Such a design is depicted in Figure 8.20. In general, one does not take into consideration that the different measurements may have a mutual influence on each other. If, and there may be good reasons for this, the dependent variables are recorded for each subject and each experimental condition in the same timely order, e.g., in the sequence blood pressure, heart rate, and state of health, nothing can be said about how the treatment affects the heart rate or the state of health. Conclusions are only possible with respect to 1. The effect of the treatment on blood pressure, 2. The effect of the treatment on heart rate after measuring blood pressure, and 3. The effect of the treatment on state of health after first blood pressure and second heart rate were measured.

Group 1	T	M_P	M_Q	M_R
Group 2	C	M_P	M_Q	M_R

Figure 8.20: Two-group design used to establish the effect of a treatment (T) in comparison with a control condition (C) by measuring three dependent variables (P, Q, and R) which are recorded in this order

If the effect of a treatment on the single dependent variables is to be established without being blurred by a possible influence of the corresponding other dependent variables, the more large-scale design in Figure 8.21 could be used. By comparing group 1 and 2 or group 3 and 4 or group 5 and 6, respectively, information is obtained about an effect of the treatment on the dependent variable P or Q or R, respectively. Only in rare cases one may be interested to know which effect the timely order of the measurements has. Due to the complexity of this problem we depict in Figure 8.22 a corresponding design which is restricted to only two dependent variables P and Q.

Group 1	T	M_P
Group 2	C	M_P
Group 3	T	M_Q
Group 4	C	M_Q
Group 5	T	M_R
Group 6	C	M_R

Figure 8.21: Six-group design for establishing the treatment effect of a treatment (T) in comparison with a control condition (C) by the measurement of three dependent variables (P, Q and R)

By comparing M_{P1} with $M^{(2)}_{P2}$ or $M^{(1)}_{P1}$ with $M^{(3)}_{P2}$, respectively, one establishes whether a measurement of P is influenced by a preceding measurement of Q. If both comparisons yield different outcomes, the extent of the influence of Q depends on whether the treatment or the control condition was used. By comparing M_{Q1} with $M^{(2)}_{Q2}$ or $M^{(1)}_{Q1}$ with $M^{(3)}_{Q2}$, respectively, it is established whether the measurement of variable P influences a succeeding measurement of variable Q. Again the effect can depend on whether the treatment or the control condition was used.

Group 1	T	M_{P1}	
Group 2	C	$M^{(1)}_{P1}$	
Group 3	T	M_{Q1}	
Group 4	C	$M^{(1)}_{Q1}$	
Group 5	T	$M^{(2)}_{P1}$	$M^{(2)}_{Q2}$
Group 6	C	$M^{(3)}_{P1}$	$M^{(3)}_{Q2}$
Group 7	T	$M^{(2)}_{Q1}$	$M^{(2)}_{P2}$
Group 8	C	$M^{(3)}_{Q1}$	$M^{(3)}_{P2}$

Figure 8.22: Eight-group design used to test the influence of a preceding dependent variable (P or Q) on a subsequent one after the application of a treatment (T) or a control condition (C)

8.2 Designs with more than One Independent Variable

Assume that a vitamin deficiency due to malnutrition was found in children in a United Nations refugee camp. The children are to receive a vitamin compound containing the vitamins A and C as a remedy. Since there are contradictory opinions with respect to the effectiveness of such a compound and with respect to the optimal daily doses, first a small sample of 18 children is randomly selected to which compounds of different compositions are applied. For vitamin A the daily doses of 0 mg, 2 mg, and 4 mg are considered and for vitamin C the daily doses of 0 mg, 60 mg, and 120 mg. The judgement of a doctor about the children's general state of health, which is given via a rating scale extending from 1 (very bad) to 10 (very good), serves as a dependent variable. This yields the two-factor design in Figure 8.23.

	Vitamin C		
Vitamin A	0 mg	60 mg	120 mg
0 mg	0 / 0	0 / 60	0 / 120
2 mg	2 / 0	2 / 60	2 / 120
4 mg	4 / 0	4 / 60	4 / 120

Figure 8.23: Two-factor design for the application of a vitamin compound with different compositions

Because the state of health varies considerably among the children, each child should "serve as its own control". It is hoped that by this proceeding the heterogeneity of the children with respect to their state of health will not have the effect that no differences can be found between the different conditions. Therefore, a repeated-measures design is used with repeated measures on the two factors "vitamin A" and "vitamin C". Each child participates for 17 weeks in the study.

In the first week each child gets one of the nine compounds daily at the same time of day. On the first day of the week the state of health of the child is rated by a doctor before the compound is given (pretest B). Another rating is performed at the first day of the second week (posttest A). From the eighth day to the 14th day, i.e. during the second week, the child gets no compound (wash-out period). In the third week a second compound is applied to the child with a rating at the 15th day (pretest) and at the 22nd day (posttest) and so on, up to the ninth and last compound, and for each compound a pretest and a posttest score is recorded (cf. Figure 8.24a, b).

Day	1	2	3	4	5	6	7
Compound	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Score	B						

Day	8	9	10	11	12	13	14
Compound							
Score	A						

Day	15	16	17	18	19	20	21
Compound	0/60	0/60	0/60	0/60	0/60	0/60	0/60
Score	B						

Day	22	23	24	25	26	27	28
Compound							
Score	A						

Day	29	30	31	32	33	34	35
Compound	0/120	0/120	0/120	0/120	0/120	0/120	0/120
Score	B						

Day	36	37	38	39	40	41	42
Compound							
Score	A						

Day	43	44	45	46	47	48	49
Compound	2/0	2/0	2/0	2/0	2/0	2/0	2/0
Score	B						

Day	50	51	52	53	54	55	56
Compound							
Score	A						

Figure 8.24a: Lapse of time of a systematic two-factor repeated-measures design (week 1-8) with repeated measures in both factors with the nine level combinations of Figure 8.23 and with pretests (B) and posttests (A)

Day	57	58	59	60	61	62	63
Compound	2/60	2/60	2/60	2/60	2/60	2/60	2/60
Score	B						

Day	64	65	66	67	68	69	70
Compound							
Score	A						

Day	71	72	73	74	75	76	77
Compound	2/120	2/120	2/120	2/120	2/120	2/120	2/120
Score	B						

Day	78	79	80	81	82	83	84
Compound							
Score	A						

Day	85	86	87	88	89	90	91
Compound	4/0	4/0	4/0	4/0	4/0	4/0	4/0
Score	B						

Day	92	93	94	95	96	97	98
Compound							
Score	A						

Day	99	100	101	102	103	104	105
Compound	4/60	4/60	4/60	4/60	4/60	4/60	4/60
Score	B						

Day	106	107	108	109	110	111	112
Compound							
Score	A						

Day	113	114	115	116	117	118	119
Compound	4/120	4/120	4/120	4/120	4/120	4/120	4/120
Score	B						

Day	120						
Compound							
Score	A						

Figure 8.24b: Lapse of time of a systematic two-factor repeated-measures design (week 9-17) with repeated measures in both factors with the nine level combinations of Figure 8.23 and with pretests (B) and posttests (A)

The design in Figure 8.24a, b has some obvious disadvantages. First, one cannot expect that in a study lasting 120 days all children will be available for the whole duration (**experimental mortality**, cf. Section 3.2.7). Further, one had better not expect that for all subjects all 18 scores will be available and that the complex design can be performed as scheduled for all subjects. It is not very likely that the doctor is the same for the whole period, which means that measures of different reliability and calibration may result. In order to avoid a systematic rater bias a strict blinding (cf. Section 4.9) should be used, i.e. not only the rating doctor and the subjects but also the doctor who applies the compounds to the children should not know the composition of the compounds. In particular, the compounds should not be applied in a systematic way over time with the low-dose compounds at the beginning and the high-dose compounds at the end of the study as displayed in Figure 8.24a, b. If this procedure is being used the effect of a compound will be inseparably confounded with the effects of a general trend in time, i.e. evident differences between the effects of different compounds are no longer necessarily caused by the composition of the compounds. It seems natural that, e.g., by a better nourishment of the children the vitamin deficiency will be overcome in the course of time and the general state of health will be improved.

Therefore, one had better assign the nine compounds randomly to the nine corresponding weeks. Still it would be better to perform this random assignment separately for each child. Since we started with a sample of 18 children we might also consider an incomplete counterbalancing (cf. Section 4.8). For this we would select 18 different sequences of the nine compounds such that each compound occurs exactly two times in each of the nine weeks. I.e., for instance the first compound occurs for two children in the first week, for two other children in the second week, etc. The 18 sequences are randomly assigned to the 18 children.

In spite of the wash-out periods one cannot rule out that a compound interacts with compounds, which are given after this compound. This may lead to superpositions of the effects, which are not easy to control. It might be possible, e.g., that after a high dose of a vitamin the need of this vitamin is met for weeks and, therefore, the succeeding compounds can have no effects. This problem is probably not so urgent in case of vitamin C, which is water soluble but may be present for vitamin A, which is fat soluble (possible forming of depots, even danger of a hypervitaminosis). This kind of carry-over effects cannot be eliminated by means of statistical adjustments, without any quite implausible assumptions. In order to measure such after-effects of compounds a far more large-scale design using the idea displayed in Figure 8.16 should be used. However, if the measurement of after-effects is not of interest, we could consider only the first eight days of the design in Figure 8.24a in the incomplete counterbalanced design with 18 sequences described above, and skip not only the rest of the design but also the pretest on day 1. Then a simple two-factor design as in Figure 8.23 would result without repeated measures and with two subjects for each level combination.

An intermediate stage between the two-factor design with repeated measures for both factors and the two-factor design without repeated measures is the two-factor design with repeated measures for only one factor. In our example, one can randomly split up the sample with 18 children into three groups each with six children. These groups are randomly assigned to the three levels "0 mg vitamin C", "60 mg vitamin C", and "120 mg vitamin C". Now it would be possible to use a completely balanced design for each of the three groups for the three levels of the second factor ("0 mg vitamin A", "2 mg vitamin A", and "4 mg vitamin A"). This could be done by randomly assigning the six possible sequences (0 mg / 2 mg / 4 mg, 0 mg / 4 mg / 2 mg, 2 mg / 0 mg / 4 mg, 2 mg / 4 mg / 0 mg, 4 mg / 0 mg /2 mg, and 4 mg / 2 mg / 0 mg) to the six children within a group. This repeated-measures design requires each child to participate in the study for 36 days, thus causing similar interpretational problems as the design with repeated measures for both factors.

In Section 7.2 and 7.3 we have seen that the interpretation of the simultaneous effects of several independent variables might be complicated because of possible interactions. This problem is aggravated if we admit not only several independent variables but also repeated measures. As we have already seen, most causal relations, which are of interest, can be detected without problems by using one-factor designs without repeated measures. In our example above only a few people will be interested in finding out the effect of certain dose combinations on the following dose combinations. Thus, a repeated-measures design is actually not needed. If one only wants to know which dose of vitamin A or C, respectively, should be applied, the one-factor design without repeated measures depicted in Figure 8.25 should be used. Here, the 18 children must be randomly split up and assigned to the five independent groups, e.g., by using the subsample sizes 4, 4, 4, 3, and 3. If, in practice, both vitamins are to be applied simultaneously and one cannot rule out that the effects of both drugs may mutually strengthen or weaken each other, the design in Figure 8.23 should be preferred to the one in Figure 8.25. Here, the children must be randomly split up into groups such that two children are assigned to each of the nine dose combinations in Figure 8.23. The course of time should be the same for each child and should be chosen analogously to the one in Figure 8.25. If one wants to study which effects the repeated application of a vitamin compound has, the design in Figure 8.15 should be used for each of the nine dose combinations in Figure 8.23.

	Day	1	2	3	4	5	6	7	8
Group 1	Placebo (0 mg)	0	0	0	0	0	0	0	
	Score								A

	Day	1	2	3	4	5	6	7	8
Group 2	Vitamin A (2mg)	2	2	2	2	2	2	2	
	Score								A

	Day	1	2	3	4	5	6	7	8
Group 3	Vitamin A (4mg)	4	4	4	4	4	4	4	
	Score								A

	Day	1	2	3	4	5	6	7	8
Group 4	Vitamin C (60 mg)	60	60	60	60	60	60	60	
	Score								A

	Day	1	2	3	4	5	6	7	8
Group 5	Vitamin C (120 mg)	120	120	120	120	120	120	120	
	Score								A

Figure 8.25: One-factor design for the application of vitamin A or C, respectively, with a posttest (A)

Summary

1. Designs with one or more factors where repeated measures are used for each subject do, in general, not permit a causal interpretation and should, therefore, not be used.

2. If it is not actually the object to investigate the influence of preceding measurements or treatments, respectively, it is possible to use designs, which permit a causal interpretation instead of repeated-measures designs.

3. If, indeed, after-effects of measurements or treatments, respectively, are the object of a study, the original sample should be randomly split up into independent groups. For each effect of interest a pair of independent groups should be formed in which the two groups differ only at one point with respect to the arrangement of measurements or treatments, respectively.

Questions

8.1. Give examples for situations where repeated-measures designs are actually needed.

8.2. When is it appropriate to perform more than one posttest after a treatment?

8.3. Explain by means of an example why the number and the pattern of pretests and posttests in Figure 8.7 and 8.8 should be kept constant for both groups.

8.4. Describe situations where pretests should be used.

8.5. Explain according to which aspects the duration of periods without treatment or control condition, respectively, and without measurement should be chosen in repeated-measures designs.

8.6. Explain the importance of experimental mortality for repeated-measures designs and for designs without repeated measures.

8.7. Which interpretational problems occur when the repeated-measures design in Figure 8.13 is being used?

8.8. Extend the four-group design of Solomon in Figure 8.19 to the case with two treatments and explain the interpretation of the outcomes.

8.9. Why is it no convincing argument to emphasize the use of each subject "as its own control" in repeated-measures designs as an exceptional advantage?

8.10. What does it mean that in studies of long duration the reliability and the calibration of measures might be impaired?

8.11. Depict a design for 18 subjects based on incomplete counterbalancing as an alternative to the design in Figure 8.24a, b.

8.12. How many groups have to be used in a design, which would permit, by analogy to the design in Figure 8.16, to control arbitrary transfer effects to succeeding measurements for the situation considered in Figure 8.24a, b?

8.13. Depict explicitly the two-factor design with repeated measures for only one factor mentioned in Section 8.2.

9 Single-Case Experimental Designs

9.1 Basic Principles of Single-Case Experimental Designs

As already discussed repeatedly, the last time in Chapter 8, considerable problems with respect to causal interpretations of the outcomes are typical of designs with repeated measures. One has to expect that these problems are particularly grave if only one sample is used and, even more, if this sample consists of only one subject. In general the outcomes of such single-case studies cannot be causally interpreted.

There might be situations, however, where samples with more than one subject are not at hand or where there are well-founded reasons that corresponding subjects will respond differently with respect to the respective independent variable. Consider a patient as a first example, whose illness has a long individual history and who, therefore, cannot be compared to other patients with respect to the choice of the optimal therapy. A population of patients with the same diagnosis might be a second example, if only one specific treatment of several possible treatments is believed to be the optimal for each single patient though one does not know which therapy should be chosen for a particular patient. A study where the existence of such individually optimal treatments was to be proved in the case of anxiolytics has been reported by Wurthmann et al. (1996).

Most proposals for the design of single-case studies cannot be considered suitable for drawing causal conclusions for two reasons. First, one cannot rule out that time trends, transfer effects etc. can yield possible alternative explanations for apparent effects of independent variables on a dependent variable just as in case of repeated-measures designs for groups. Second, the existence of complicated dependence structures for successive measurements cannot be ruled out for repeated measures, with the consequence that assumptions are required for the statistical evaluation whose validity is by no means assured. This criticism concerns all kinds of time-series analyses. One should not believe that the validity of these assumptions can be checked statistically, since the customary statistical procedures only allow to prove the invalidity of assumptions. If the result of a statistical check does not indicate a violation of the assumptions in question, this does not mean at all that the assumptions are valid (cf. Section 3.1.1). Therefore, the interpretation of the results of time-series analyses will be always problematic.

However, there exists one approach to the planning of single-case experiments which permits to conclude that an effect of the independent variables is present in case of a statistical significant result, if the experiment and likewise its evaluation have been appropriately performed. What one cannot really rule out, however, is that the observed effect has been affected in a way different from the assumed.

The starting-point for this kind of single-case experiments is a fictitious experiment by Sir Ronald Fisher which has already been described in Section 1.7. Here, eight cups of tea were prepared. In four of these first tea (T) and then milk, in the remaining four first milk (M) and then tea was poured. A lady was to find those four cups of tea, in which tea had been poured first into (case T), by merely tasting the tea. For this, the eight cups of tea were presented to the lady in a random arrangement.

Altogether, there are 70 ways in which four M-cups and four T-cups can be arranged. This can be seen in the following way: the first M-cup can be placed in eight different ways on eight saucers. If this cup is placed, there remain seven ways to place the second M-cup. For the third M-cup are left six ways, and for the fourth M-cup five ways for placing the cup. Altogether these are 8 × 7 × 6 × 5 possible different arrangements. However, in this kind of reasoning the four M-cups are numbered, i.e. that means that, e.g., the arrangement "First M-cup on place 1, second M-cup on place 5, third M-cup on place 6, fourth M-cup on place 8" is considered as different from the arrangement "Second M-cup on place 1, first M-cup on place 5, third M-cup on place 6, fourth M-cup on place 8". The four integers 1, 2, 3, and 4 can be assigned to the four places in 4 × 3 × 2 × 1 = 24 ways: the integer 1 can be assigned in 4 ways, after this the integer 2 in 3 ways, after this the integer 3 in 2 ways, and after this the integer 4 in 1 way. Because those arrangements of the M-cups, which are different solely due to the arbitrary numbering of the cups, should not be distinguishable by the lady, we observe that in the above given number of arrangements, each arrangement occurs 24 times. From this follows that there are only (8 × 7 × 6 × 5) / 24 = 70 essentially different arrangements of the M-cups. Because the four saucers, to which no M-cup is assigned are automatically occupied by T-cups, there are just 70 different ways to arrange four M-cups and four T-cups.

The lady decides for one of these arrangements that it is the correct one. The probability of finding the correct arrangement simply by guessing is given by (1 / 70) = .014. If this probability is regarded as small, e.g., because it does not exceed the customary reference value of .05, a significant result is assumed if the lady announces the correct arrangement.

If a perfect outcome is not achieved, the second-best outcome is given by an arrangement where exactly one M-cup is falsely identified as a T-cup and, consequently, one T-cup as an M-cup. The four ways to falsely identify an M-cup as a T-cup have to be combined with the four ways to falsely identify a T-cup as an M-cup. There are 4 × 4 = 16 such combinations. If the competence of the lady would already be recognized if she identified correctly at least six of the eight cups, we can compute the probability of identifying correctly eight or six cups solely by guessing as $\frac{1}{70} + \frac{16}{70} = .243$. This probability is so high in comparison with the reference value of .05 that if only six cups are correctly identified, doubts should arise with respect to the ability of the lady to identify the way in which the tea was prepared by tasting it.

For good reasons, Senn (1994, p. 223) pointed out that the considerations above are only valid if the following two conditions are fulfilled: First, an **open protocol** has to be assumed, i.e. the lady has to know that exactly four cups of each kind are presented to her, i.e. that, e.g., an answer with five cups of one kind and three cups of the other kind is not permitted. Second, the lady should know that a random arrangement is used, where each of the possible arrangements occurs with the same probability. This holds likewise, e.g., for the two arrangements MMMMTTTT or TTTTMMMM.

If the color of the tea differs for the two kinds of preparing it, the lady may present the correct identification though there do not exist differences in taste or though the lady is not able to perceive such differences in taste if they exist. This means that an effect is found which is caused by the independent variable, i.e. by the way the tea is prepared. However, the apparent causal relation that it is the taste of the tea which permits the correct identification of the way of preparing the tea does not exist in reality. Therefore, for this fictitious experiment one has to require, as was also pointed out by Senn (1994) that an absolutely working **blinding** (cf. Section 4.9) has been introduced.

One might want to try to guarantee such a blinding as follows: the tea is not poured into cups but into non-transparent bottles, which do not differ optically or by material. Four of the bottles are randomly selected and the tea with the preparation M is poured into them. The remaining bottles are filled with tea of the preparation T. The eight bottles are filled in a random order to avoid, e.g., systematic differences in temperature. Then the eight bottles are placed into a box with warm water to equalize completely possible differences in temperature. Finally, the bottles are brought into a random arrangement and are presented to the lady by a subject who has neither observed the filling of the bottles nor their arrangement. The lady has to drink the tea directly from the bottles.

Obviously, the fictitious experiment described is a single-case experiment. As it has not been asked whether it is possible, in principle, whether the way of preparing the tea can be detected from its taste, but rather, whether this particular lady has this ability, a single-case experiment must be used as a group experiment could not help to answer this question.

For this experiment as for each design with repeated measures many influences can make it difficult to give a causal interpretation of the outcomes: one reason could be timely trends which might have the effect that the conditions for the first trials differ from those for the following trials. This can be caused, e.g., by a gradual neutralization of the taste buds or by lower temperatures for later trials. A further reason can be asymmetric transfers because, e.g., the first occurrence of an M-trial may have the effect that the lady cannot discriminate the following M- and T-samples from the first M-sample. Nevertheless, the lady would pay attention to "identify" exactly four M-samples.

All these influences might have the consequence that the probability that an existing causal relation is not found increases. If the correct identification is presented and if the blinding has been successful, it can be argued that the lady can discriminate the two ways of preparing the tea just by tasting the tea, and the probability that the correct result is due to chance is given by (1 / 70).

Except for practical considerations it might have been better to use a larger number of samples in Fisher's fictitious experiment, e.g., 12 samples with 6 M-samples and 6 T-samples. In this case $(12 \times 11 \times 10 \times 9 \times 8 \times 7) / (6 \times 5 \times 4 \times 3 \times 2 \times 1) = 924$ arrangements would have been possible. Here, we would have had one totally correct and $6 \times 6 = 36$ arrangements with only one confusion of a M- with a T-sample. The probability to find a totally correct arrangement or one with only one confusion by chance alone, would be given by $\frac{1}{924} + \frac{36}{924} = .040$. In this design one might tolerate one confusion of cups for drawing the causal conclusion that the lady can identify the way of preparing the tea by tasting it.

Of course, Fisher's fictitious experiment is of no great interest with respect to its practical relevance. Far more important is the structure of the design of this experiment, which permits to conceive single-case experiments permitting a causal conclusion. On this basis many proposals for single-case experiments were made by Eugene S. Edgington in Chapter 12 of his book (Edgington, 1995) and in many articles (Edgington, 1967, 1972b, 1975, 1980a, b, c, 1982, 1984, 1987, 1992, 1996; Edgington and Bland, 1993). Further it was described in which ways these experiments could be evaluated by means of distribution-free randomization tests. It should be pointed out that for this kind of experiments solely an appropriate distribution-free analysis can be used because one has to expect that the scores measured at one subject (repeated measures) are dependent.

9.2 Selected Single-Case Experimental Designs

Here, the construction of experimental single-case designs is illustrated by two examples. Further applications can be found in the book and the articles by E. S. Edgington.

9.2.1 Design with only One Factor with only Two Levels

We, again, consider the example in Section 8.1 where it was investigated whether the intake of an anorectic yields a reduction of weight. In particular, we consider once again the design in Figure 8.11. We simplify this design by omitting the pretests B and apply the resulting design to a single subject as depicted in Figure 9.1.

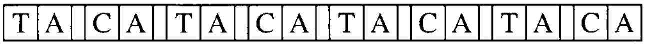

Figure 9.1: Single-case design with treatment (T, four times), control (C, four times), posttest (A, eight times), and periods without treatment or measurement (seven times)

As already discussed in Section 8.1, the experiment should be performed as a double-blind study. This means that the subject does not know whether he or she is exposed to the treatment or control condition at a given moment and the experimenter does not know it either. In order to facilitate the detection of a treatment effect, the duration of the periods without treatment or measurement should be chosen so large that no physiological effect of the preceding treatment or control condition can be expected. Here, irreversible effects, e.g., produced by forming depots or other kinds of effect accumulation cannot be ruled out. Furthermore, in addition to the physiological effects psychological after-effects are also conceivable. Finally, the requirement that treatment and control cannot be discriminated by subjects and experimenters cannot really be met if there are side-effects of the treatment.

In Section 8.1 we have already pointed out that the systematic alternating arrangement of treatment and control as in Figure 9.1 has the consequence that several plausible alternative explanations are at hand if an apparent effect results. In an experimental single-case design one of the 70 possible arrangements of treatment and control is selected by chance, i.e. with the probability (1 / 70). The subject will be exposed to this arrangement and the eight weight scores (A) are recorded. If the four smallest scores occur for the four treatment conditions a weight reducing effect of the anorectic will be assumed with an error probability of (1 / 70) = .014. Note that the number of possible arrangements (70) is derived in analogy to the fictitious experiment by Fisher (cf. Section 9.1).

As opposed to the consideration above, it is not necessary for this problem that the four smallest weight scores occur for the four treatment conditions. Rather, it would have been possible to compute, e.g., the sum of weight scores for the corresponding four treatment conditions for each of the 70 possible arrangements. If the measurements are performed with sufficient precision it is highly probable that the 70 resulting sums are all different. The probability that one of the three smallest weight sums is obtained by chance is then given by (3 / 70) = .043. Therefore, if one of these three smallest weight sums has been found for the randomly selected arrangement, a weight reducing effect of the anorectic can be assumed, at least with an error probability of .043. This procedure corresponds to that of the randomization tests for single-case experiments proposed by E. S. Edgington.

9.2.2 Interrupted Time-Series Design

If an irreversible effect of the drug cannot be ruled out it will not be possible to detect a weight-reducing effect of the anorectic by the design above. In such a case a time-series design as in Figure 9.2 might be more appropriate. Here, in addition to the treatment (T) a control condition (C) is used. Both conditions are applied in a double-blind study, i.e., the subject and the treating doctor are not able to discriminate the two conditions. The day of intervention is randomly selected from the days 11 to 110, i.e. the intervention occurs at the earliest at day 11 and at the latest on day 110, though neither the subject nor the treating doctor know this. The first 10 days are considered as a kind of habituation phase. Since the subjects know that the duration of the trial is 120 days it is better to use the last days as a fading-out phase. We use the sum of the weight scores on the intervention day and on the following four days as a measure of the treatment effect. As the intervention day was randomly selected from altogether 100 days, the probability that the smallest sum of weight scores, i.e. also the five smallest weight scores, is obtained equals (1 / 100) = .01 if the effect of the drug does not differ from that of the control condition, and if the resulting 100 sums are all different. If one assumes that the anorectic has no effect, the probability of obtaining one of the five smallest weight sums for the randomly selected intervention day is given by (5 / 100) = .05. Thus, if one of these sums results for the selected arrangement, one might claim that the anorectic has a weight-reducing effect if an error probability of .05 is accepted.

Day	1	2	...	30	31	32	...	120
Condition	C	C		C	T	C		C
Score	A	A		A	A	A		A

Figure 9.2: Lapse of time of an interrupted time-series design over 120 days with an intervention (T) at a randomly selected day (here: day 31), a control condition (C), and posttests (A)

9.3 An Alternative Principle of Single-Case Experimental Designs

A possible modification of Fisher's fictitious experiment (cf. Section 9.1) might be to not ask the lady to identify four M-cups and four T-cups but to identify the tea-preparation method for each of eight cups, where for each cup it is randomly decided whether it is to be an M-cup or a T-cup. This yields $2^8 = 256$ possible different arrangements of the cups altogether, each of which occurs with the same probability (1 / 256) = .004. This is because we have two possibilities for the first cup, which are to be combined with the two possibilities for the second cup. This yields 2 × 2 possibilities which, together with the two possibilities for the third cup, yield 2 × 2 × 2 different outcomes, etc. Thus, the probability that the lady finds the correct result by guessing is considerably smaller here than in Fisher's proceeding where the numbers of M-cups and T-cups were fixed in advance and where the chance probability was equal to (1 / 70).

If we apply this alternative principle to the anorectic example in Section 9.2.1, we might compute the mean of the weight scores for the treatment days for each of the 256 possible arrangements. For the specific arrangement, in which the control condition is applied for all days, the mean would be set to zero. If the mean of the weight scores for the treatment days in the actually realized arrangement is one of the 12 smallest of the 256 means, a weight reducing effect of the anorectic might be assumed with an error probability of (12 / 256) = .047. Here, the 256 resulting means of the weight scores are assumed to be different from each other.

From the viewpoint of experimental design note that not only such difficult to interpret arrangements as TTTTCCCC and CCCCTTTT may occur just as for the Edgington approach but in addition arrangements of the form TTTTTTTT or CCCCCCCC. In particular, the last arrangement would always yield a treatment effect contradicting intuition. These might be reasons why this alternative proposal has not been considered yet.

9.4 Combination of the Results of Several Independent Single-Case Experimental Designs

Sometimes an independent single-case experiment is performed for each subject in a sample and one would like to generalize the results by pooling them over the subjects of the sample. One such method was also described by Eugene S. Edgington in another context (Edgington, 1972a). This method is here explained though many authors, e.g., R. A. Fisher, have proposed other procedures for the same purpose.

The first step is to compute a P-value for each subject. If the sixth-smallest sum of weight scores from altogether 70 possible sums in Section 9.2.1 has been assigned to a subject, the corresponding P-value is given by (6 / 70) = .086. This P-value is the probability to obtain the sixth-smallest or an even smaller sum of weight scores under the assumption that the anorectic has no specific effect. A very small P-value, e.g., a P-value smaller than .05, indicates that presumably an effect of the treatment is present for the corresponding subject. A large P-value is an indication that no treatment effect exists. In our example, the 70 P-values (1 / 70), (2 / 70), ..., (70 / 70) are possible. If the anorectic has no effect, each of the 70 possible P-values occurs with the same probability.

Assume now that the different subjects, for which single-case experiments are performed, do not mutually influence each other. If this assumption is not justified, the procedure described in the following must not be used.

If the anorectic has not had an effect, the probability for a subject to obtain one of the three smallest weight scores is given by (3 / 70) = .043. This is, at the same time, the probability to obtain one of the three largest weight scores. If there is no effect of the anorectic, large and small P-values occur with the same probability. Thus, the probability that several independent subjects at the same time have only small P-values and none has a large P-value will be small if the anorectic does not have an effect.

If the P-values (P_1, ..., P_k) are obtained for k subjects, which do not influence each other, according to Edgington (1972a), the sum of these P-values is computed:

$S = P_1 + ... + P_k$.

The probability (P_T), to obtain this sum (S) or a smaller value, if the anorectic has had no effect, is computed in the following way:

First, the largest integer which is smaller than or equal to S is found. This integer is denoted by $[S]$. Then with the definition

$$n! = 1 \times 2 \times 3 \times ... \times n, 0! = 1$$

for a natural number n we compute

$$P_T = \frac{S^k}{k!} - \frac{(S-1)^k}{1!(k-1)!} + \frac{(S-2)^k}{2!(k-2)!} - K + (-1)^{[S]} \frac{(S-[S])^k}{[S]!(k-[S])!},$$

i.e. a sum with alternating signs.

Assume, e.g., $k = 5$ subjects with the P-values $P_1 = \frac{10}{70}$, $P_2 = \frac{24}{70}$, $P_3 = \frac{33}{70}$, $P_4 = \frac{26}{70}$, and $P_5 = \frac{60}{70}$. Then we get

$$S = \frac{153}{70} = 2.186.$$

From this follows $[S] = 2$, and we get

$$P_T = \frac{2.186^5}{1 \times 2 \times 3 \times 4 \times 5} - \frac{(2.186-1)^5}{1(1 \times 2 \times 3 \times 4)} + \frac{(2.186-2)^5}{(1 \times 2)(1 \times 2 \times 3)} = .318.$$

Since this value is considerably larger than .05, no weight-reducing effect of the anorectic is assumed for that population for which the considered sample with k subjects is representative.

If the sum (S) of P-values does not exceed the value one, we have [S] = 0 and the expression for calculating P_T is reduced to its first term ($S^k / k!$).

Assume, e.g., $k = 4$, $P_1 = \frac{5}{70}$, $P_2 = \frac{10}{70}$, $P_3 = \frac{6}{70}$, and $P_4 = \frac{5}{70}$. Because of $\frac{5}{70} = .071$ each of the four P-values is larger than .05 and an effect of the anorectic could be assumed for none of the four subjects. We find

$$S = \frac{26}{70} = .371 < 1$$

and

$$P_T = \frac{.371^4}{1 \times 2 \times 3 \times 4} = .001,$$

i.e. there seems to be an effect in the population. This apparent contradiction can be explained by the observation that all four P-values are considerably smaller than .5 and that the probability for such an event is very small, if the anorectic actually has had no weight-reducing effect.

In particular, if the number (k) of P-values, which are to be combined, is very large, their sum (S) will often exceed the value of one, even if the individual P-values are small. As a consequence at least the first terms of the alternating series above will have large absolute values. If the differences of very large numbers are computed, the final result might be quite misleading due to rounding errors. It might even happen that the result is smaller than zero or larger than one though this, in theory, is not possible. Sometimes this can be avoided as follows: we do not calculate P_T for the P-values $P_1, ..., P_k$, but P_T^C for the complementary P-values $P_1^C = 1 - P_1, ..., P_k^C = 1 - P_k$. If one computes at the end $P_T = 1 - P_T^C$ one, again, gets the correct result.

As an example, we consider again the example above with $k = 5$:

Here, we obtain the complementary P-values $P_1^C = \frac{60}{70}$, $P_2^C = \frac{46}{70}$, $P_3^C = \frac{37}{70}$, $P_4^C = \frac{44}{70}$, and $P_5^C = \frac{10}{70}$. From this we derive the sum

$$S^C = k - S = \frac{197}{70} = 2.814$$

and from this

$$P_T^C = \frac{2.814^5}{1\times2\times3\times4\times5} - \frac{(2.814-1)^5}{1(1\times2\times3\times4)} + \frac{(2.814-2)^5}{(1\times2)(1\times2\times3)} = .682,\ P_T = 1 - P_T^C = .318,$$

i.e. the same value as by direct computation.

SUMMARY

1. In general, the results of single-case designs, in particular, of time-series designs, cannot be causally interpreted since potentially infinitely many alternative explanations exist due to the lack of randomization. This is obvious because single-case designs are special cases of repeated-measures designs.

2. The designs by Edgington are an exception. They are based on the idea of the fictitious experiment discussed by R. A. Fisher, where a lady has to identify two kinds of preparing tea by tasting the tea. Here, a design from a known set of designs is randomly assigned to a subject and a distribution-free evaluation is performed.

3. The results of Edgington designs permit a causal interpretation if the subjects cannot discriminate the different experimental conditions, i.e. if a true single-blind study is performed. The performance of a double-blind study, where the experimenter cannot discriminate the conditions either, is even better.

4. The results for several independent single-case experiments can be statistically combined to a total result.

Questions

9.1. Describe situations in which single-case studies are appropriate.

9.2. When is it principally not possible to interpret the outcomes of single-case studies?

9.3. What are the advantages of the single-case experiments proposed by Edgington as opposed to other single-case studies?

9.4. Which difficulties might occur if the outcomes of Edgington experiments are to be interpreted?

9.5. State, how many arrangements there are for a Fisher tea experiment with 5 M-cups and 6 T-cups, i.e. with altogether 11 cups.

9.6. Is it necessary that an open protocol exists for Edgington experiments as this is the case for the Fisher tea experiment?

9.7. In which way, in analogy to Section 9.2.1, could a one-factor design with three levels be realized in a single-case experiment?

9.8. In which way, in analogy to Section 9.2.2, is it possible to use a time-series design when an intervention is not restricted to a single day?

9.9. For $k = 12$ independent single-case experiments the P-values $P_1 = .10$, $P_2 = .30$, $P_3 = .01$, $P_4 = .60$, $P_5 = .65$, $P_6 = .20$, $P_7 = .01$, $P_8 = .03$, $P_9 = .10$, $P_{10} = .04$, $P_{11} = .02$, and $P_{12} = .01$ have been obtained. Compute P_T via the P-values and the complementary P-values.

9.10. 18 students participated in a study (cf. [9]), which was intended to investigate the effects of loperamide on anorectal functioning in healthy males. An anorectal manometry was performed for these subjects on two different days separated at least by seven days. For this aim a rectal balloon was positioned at three different locations (5, 10, or 15 cm from the anal verge). The sequence of the three locations was randomized.

One day the subjects received a dose of 10 mg loperamide, the other day a placebo, where the order of the two conditions was counterbalanced. The subjects were blind with respect to medication and the location of the balloon. The researcher who performed the manometry was blind with respect to medication. By means of an analysis of variance a "significant" main effect of the location of the balloon was found to exist with respect to a first dependent variable. With respect to a second dependent variable also a main effect of the location of the balloon existed and in addition a "significant" interaction between location and medication was found. No "significant" results were found for quite a few other statistical tests.

What can be objected against the results of the study above?

9.11. In a study (cf. [5]) one wanted to reveal whether driving performance is affected by radio monitoring, task difficulty, and time of day. The study took place with 20 subjects (12 males and 8 females) in a laboratory via a computer-controlled, simulated driving task. Each subject was exposed to eight experimental conditions, where the sequence of the conditions was randomly fixed. The subjects had to participate in the study at four days. At two of the four days the study took place between 9:00 a.m. and 11:00 a.m., at the two remaining days between 4:00 p.m. and 6:00 p.m. At one of two days on which the experiment was scheduled for the same time, a radio program was delivered via headphones, at the corresponding other day it was not delivered. In this way four different experimental situations resulted. In each of these situations an "easy" and a "difficult" task was to be performed, so that altogether eight different experimental situations were produced. A three-factor analysis of variance for repeated measures showed "significant" results for all three main effects and for all interactions with the exception of the interaction between "difficulty of the task" and "time of the day".

Discuss why this study cannot be used in order to answer the respective questions.

Answers to the Questions (with References)

Answers to Chapter 1

1.1 Randomization and the introduction of control conditions are the two most important principles of experimental design. If a researcher wants to establish that a certain cause entails a certain effect it is not enough to observe the supposed cause and the subsequent apparent effect. The researcher needs a control condition, which is equal to the experimental condition in all but one aspect namely the supposed cause. If one observed consequently that the effect occurs only if the cause is also present, one concludes that the effect is a consequence of the cause.

The main problem with this proceeding is that we can never rule out that, in real life, the two conditions might differ not only in the supposed cause but also in other aspects. In the medical and behavioral sciences this might be, e.g., differences between the observed subjects. Such aspects might on the one hand entail that the effect occurs though the supposed cause is not responsible for this or though the supposed cause is even not present. On the other hand such an aspect can invoke that the effect is not observed in the presence of the cause though the cause principally entails the effect.

Since in an empirical study the two conditions can differ in potentially infinitely many aspects, the majority of which is neither known nor measurable, deterministic causal conclusions seem to be impossible. Therefore, one is content to assure that the probability of establishing a false causal conclusion is below a given bound (α). This is achieved, e.g., by assigning the subjects at random to the conditions or, conversely, by assigning a randomly selected design to the subjects. Both kinds of random assignment are called randomization. Of course, a randomization does not guarantee that the conditions are actually totally equivalent. But it yields a control of the probability of a false causal conclusion.

If the mebership in a pre-existent population (e.g., the population of all women) is supposed as the cause of an effect, a random assignment of the subjects to both conditions (target population and rest population) is not possible, in general. If one wants to conclude that the supposed effect actually is due to the membership in the population, true random samples from the target and the rest population are needed. If the considered samples are not random samples from the respective populations it cannot be ruled out that the observed apparent effect is not due to the membership in a population but to a selection effect which has as its cause that the two samples are systematically different not only with respect to their membership in a population but also in other aspects.

1.2 The proposed design does not permit the conclusion that the better and more well-fed air of the Israelite hostages in comparison to that of courtiers of the same age is due to the different diets. It might be, e.g., that the Israelites already had a better state of health than the courtiers before the study and that ten days of vegetarian diet without alcohol could not change much. The best would have been if Melzar had randomly selected two of the four Israelites to feed them with Babylonian food while the two remaining hostages would have got the nourishment proposed by Daniel. However, this design might have met with resistance by Daniel for religious reasons. This must be assumed, in particular, if Daniel himself had been one of the two hostages, who had to eat Babylonian food due to a random decision.

Another design for which the interpretation of the outcome would not have been as easy, would have required to split up a large sample of Babylonian courtiers randomly into two subsamples which had been randomly assigned to the two diets. Even if it had been found that the sample with a vegetarian diet had had the better air, it could not have been ruled out that the reverse result might have been observed for the Israelites. However, if the Babylonian diet should have given a better outcome for the Babylonian courtiers this result could also not have been generalized without more ado to the Israelites.

1.3 In the report of Athenaeus it is not said that the assignment to the two conditions (experimental condition: a lemon is given, control condition: no lemon is given) was performed randomly after forming pairs of convicts. Otherwise, it cannot be ruled out that always the healthier convicts received the lemon, so that no causal conclusion is possible. Further, it is not permitted that convicts who survived the asps' bites were used once again in the experiment because they possibly developed a certain immunity against the venom.

1.4 Even if we are willing to believe that the outcome of the lemon experiment was as it was reported by Athenaeus, though nowadays this may be called into question, the conclusion of Democritus, who is the narrator in Athenaeus' report, is not permitted, namely that lemon is an antidote to all kinds of poisons. The outcome of the experiment is related to the venom of asps and cannot be generalized without more ado to other kinds of poisons. However, Democritus cites a statement of Theopompus of Chios whom he calls a truthful man who had invested much money in the accurate investigation of history. According to this witness a whole citron boiled in some Attic honey should protect against any kind of poison. For this, unfortunately, no experiment is reported.

It might be, of course, that the effectiveness of the lemons nowadays is not the same as in the times of Athenaeus. However, some doubt with respect to the seriousness of Athenaeus' report might arise. He says that after having heard the story of Democritus all participants of the banquet were so convinced by his words that they started to eat the available lemons at once as if they had not touched any food or drink before.

1.5 In studies with human beings or animals, environmental variables such as temperature, atmospheric humidity, and atmospheric pressure should be recorded because one cannot rule out that such variables might have influenced the observed behavior. If such variables are constant during the whole experiment the generalizability of conclusions is restricted as the found effects can maybe only be observed for certain level combinations of environmental variables. But if it is not possible to keep the environmental variables constant during the experiment one cannot rule out that the observed effects have occurred due to changes in the environment and not due to the considered causal variable.

1.6 Experiments are performed with samples of subjects. Conclusions from such experiments can be generalized only to a population, for which the respective sample is representative. Effects, which are found for a sample of young rats, do neither have to exist for a population of old rats nor for populations of pigeons, pigs or human beings. Effects, which are found for students of psychology in their first year of study, might exist only in this population.

1.7 The declaration of Mill in the text, in principle, means that one can draw a causal conclusion if an effect is observed as soon as the cause is introduced at an arbitrarily chosen point of time at an arbitrarily chosen location. In his footnote, Mill realizes the problem that a false causal conclusion is drawn if the effect is not produced by the introduced cause but by one of the means by which the supposed cause was introduced.

For example, we may randomly choose time and location where a drug is applied by which the blood pressure is to be increased. Always after applying the drug an increase of the blood pressure is observed. If the drug must be applied by means of an injection, it cannot be ruled out that not the drug but the use of an injection causes the increase of the blood pressure. To avoid such false inferences, a control condition should be used where a placebo is injected, e.g., the solvent, in which the drug was solved.

1.8 a. Method of Agreement

We study the way of life of a sample of patients, which complain of a hitherto unknown disease of the respiratory ducts. It is found that the patients seem to exhibit no common characteristics and differ with respect to age, gender, educational standard, income, employment, housing conditions etc. The only common characteristic could be that all patients have laid carpets in their homes. Because it is not plausible to assume that all patients laid the carpets instead of using an alternative because of their disease, e.g., because their doctors recommended this, and because this possibility can also be ruled out by an interview, the carpets can be only a cause but not an effect of the disease. However, it cannot be ruled out that other common characteristics were overlooked. Further, the real cause of the disease may be the glue by which the carpets were fixed to the ground or a certain impregnating agent for carpets.

b. Method of Difference

A man is found among the patients in the example above who has a twin brother who has not the disease though he works in the same department of the same enterprise, lives in the same house and has also a carpet in his rooms. However, both carpets were manufactured by different firms. Because no other differences can be found in the modes of life of the two brothers it seems reasonable to assume that the carpet produced by the one firm is responsible for the disease or at least is connected with the cause, if it can be ruled out that this carpet was bought as a consequence of the disease.

c. Indirect Method of Difference

In the example above the way of life has been studied for samples of sick and healthy persons and no differences were found with respect to age, gender, educational standard, income, employment, housing conditions etc. It was solely observed that all patients had carpets, produced by a certain firm *X*, laid in their homes while this was not the case for all healthy persons. This is a reason for assuming that the carpets of firm *X* either cause the disease or are connected with this cause. The other possibility that the patients bought the carpets of firm *X* because of their disease can be ruled out by an interview.

d. Method of Residues

In two villages doctors observe for several residents a common disease syndrome. Patients with this syndrome show eczema, have problems with their respiratory ducts and, in addition, complain about colics. One finds out that all patients are working in a local pharmaceutical factory, that birch-trees grow at the main streets of both villages, and that the houses of all patients are not connected to the public water supply but have private wells. A further investigation reveals that other employees of the factory who do not live in the two villages also have problems with their respiratory ducts though they do not exhibit the two other symptoms. Further it is observed that several residents of the two villages who do not work in the pharmaceutical factory and whose houses are connected to the public water supply have eczema but no problems with their respiratory ducts and no colics. The conclusion is drawn that the cause for the problems with the respiratory ducts is related to the pharmaceutical factory and that the eczema might be due to an allergy caused by the birch-trees. After this it seems established that there remains as a cause for the colics only polluted ground water.

That this kind of reasoning is not really conclusive is obvious because there may be still other aspects in which the way of life of the patients may differ from that of other people. This is also conceded by Mill (1846, pp. 464-465):

> "As one of the forms of the Method of Difference, the Method of Residues partakes of its rigorous certainty, provided the previous inductions, those which gave the effects of A and B, were obtained by the same infallible method, and provided we are certain that C is the *only* antecedent to which the residual phenomenon c can be referred; the only agent of which we had not already calculated and subducted the effect. But as we can never be quite certain of this, the evidence derived from the Method of Residues is not complete, unless we can obtain C artificially and try it separately, or unless its agency, when once suggested, can be accounted for, and proved deductively, from known laws."

In the context of our example the proposals of Mill mean that either water from the private wells should be given to subjects without colics, in order to find out whether this leads to colics. Alternatively water of the private wells should be analyzed with respect to agents or organisms which are known to cause colics.

e. Method of Concomitant Variations

In a rural area each spring one observes that the eyes of many people run with tears, that their nasal mucosa are irritated, and that some suffer from an itching skin. These effects do not occur exactly at the same time each year but sometimes earlier and sometimes later. Finally, it is found that the symptoms are always observed approximately one day after the pollen of stone-pines is set free and subside a short time after the day where no more pollen can be observed. Because it is not plausible that the pollen is set free as a consequence of the symptoms it is concluded that the pollen causes the symptoms.

However, this way of reasoning is questionable in more than one aspect. First, only an association is observed between the occurrence of the pollen and the symptoms. This does not permit a causal conclusion because a third variable may exist, e.g., certain weather conditions, by which at the same time the pollen is set free and the symptoms are caused, though the pollen itself has no influence on organisms. Second, not everybody who is exposed to the pollen exhibits the symptoms. Thus, if the pollen actually is responsible for the symptoms, this is true only if at the same time in addition one or more other causal variables exhibit certain levels.

1.9 By "arbitrary" Wundt (1911, p. 25) means that the experimenter, in principle, is free to decide when an experiment takes place, at which points of time observations are made, which variables are observed, which interventions by the experimenter are performed and at which points of time these interventions are performed. This arbitrariness is not given if, e.g., it is to be investigated which consequences a legal ban on smoking in public has on smoking behavior in general. As a rule the researcher, in this case, is neither able to fix the point of time where the ban passes into law nor can he or she exercise an influence on the regulations of the law.

Wundt does not mean by "arbitrariness" that the experimenter is completely free in the choice of the methods and timing. If, e.g., it is to be tested in an animal study whether a drug may improve memory, such an experiment would be meaningless if the experimenter would apply so high doses of the drug to the animals that these would immediately die. Also a measurement of memory performance just at the moment where the drug is applied in many cases would not be advisable.

1.10 To investigate whether a drug enhances memory performance of rats, each of in total 20 rats is set alone in a water maze. This is a circular basin filled with warmed water in which at a certain place under the surface of the water but near to it a platform is positioned which can be seen by the rats. The rats learn to swim to the platform where they can take a rest. Thereafter, the sample of rats is randomly split up into two subsamples with 10 rats each and the rats of one subsample get the drug, the rats of the other subsample a placebo, e.g., the solvent used to solve the drug. A week later the rats again are set into the water maze, where, however, this time the water has been made opaque by adding milk. For each rat the time needed for finding the platform is recorded. If the rats in the drug group exhibit shorter time scores than the rats in the placebo group, a better memory performance of these rats is assumed.

A possible objection to this interpretation of the outcomes might be that the drug possibly raises only the activity of the rats but by no means improves memory performance. Then the rats receiving the drug would find the platform early for the sole reason that they covered a longer distance in the water maze in the same time. Therefore, by this kind of experimental design it is not possible to isolate memory performance from activity.

At least the above objection of a possibly enlarged activity might be avoided by another memory paradigm. Here, rats learn that they get only food if they press in a Skinner box first once the left lever, then twice the right lever, then twice the left lever, and finally once the right lever. Then the rats receive drug and placebo as in the experiment above. After a week the time is measured the rats need for performing the task. Of course, an increased activity caused by the drug would have the effect that the drug rats exhibit a higher frequency of lever pressing than the placebo rats. However, it is not very plausible that the rats would show the learned lever-pressing pattern with a high probability if only activity but not memory was influenced by the drug.

Unfortunately, also for this paradigm the isolation of activity and memory performance is not completely successful, because it cannot be ruled out that the drug enlarges at the same time activity and memory performance. Further, if the drug influences only activity but not memory, and this has as a consequence a higher lever-pressing rate of the drug rats, for these rats the probability is higher that they exhibit the learned pattern in a given time by pure chance.

If one uses as a dependent variable the time up to the first reinforcement or, alternatively, the number of correct solutions of the task in a given time interval, again it cannot be ruled out that the drug has solely increased activity but not memory performance. This is seen, if we assume that as well under drug as also under placebo a certain rest of memory is left. By an increased activity of the drug group in combination with this rest of memory the required lever-pressing pattern would be shown earlier and with a higher frequency. A better dependent variable might be the number of failing attempts before the first mastering of the task.

1.11 The authors themselves quote as the decisive advantage of randomization that a subject cannot predict which condition will be presented to him or her next. By this it is rendered more difficult, among other things, that preceding stimuli are used as reference points for succeeding stimuli. But even more important seem to be the advantages which result from the fact that effects by memory, exercise, fatigue etc. are not connected inseparately with certain presented sequences of stimuli.

1.12 a. Randomization means that a sample of subjects is randomly split up into subsamples and each of these subsamples is randomly assigned to another level of a causal variable or to another level combination of causal variables, respectively.

b. Randomization means that a design is randomly chosen from a given set of designs and the chosen design is assigned to a subject or to a sample of subjects, respectively.

c. Randomization means that true random samples of a given size are drawn from two or more populations. Here, we have a true random sample from a population if each other sample of the same size has the same probability to be selected.

Hence the notion of randomization is used for very different proceedings. In each case the object is to rule out the possibility of systematic biases of experimental situations by introducing an additional random experiment. Such biases might have been caused by known or unknown extraneous variables and have the consequence that alternative explanations can be formulated for the outcome of an experiment.

1.13 a. The outcome of a randomization procedure may be unsatisfactory from the viewpoint of a deterministic control of extraneous variables though the probability for the assertion of a false causal conclusion is controlled.

If, e.g., the effectiveness of a supposed anorectic is to be proved, the original sample of 20 subjects is randomly split up into two subsamples and one of the two subsamples is randomly assigned to the supposed anorectic, the other sample to a placebo. If, in this random partition, the ten subjects with the lowest weight scores are assigned to the anorectic group we have a problem with the interpretation of the outcome. It might be difficult to decide whether this group has lower weight scores after the treatment than the placebo group, due to the anorectic or because this group has already had lower weight scores before the treatment.

b. If a researcher claims to have performed an appropriate randomization and, hence, has controlled all extraneous variables in a statistical sense this is difficult to check afterwards. One cannot check either whether the researcher used an unsuitable, not really random procedure, e.g., by assigning patients entering a study in a systematic way alternatively to the different experimental conditions depending on the date of their entrance. Or, no randomization has been used at all, but, e.g., the first ten patients were assigned to an experimental group and the next ten patients to a control group. Or, a systematic selection took place where seriously ill patients received a standard treatment while the other patients were assigned to a new therapy.

c. A random partition of subjects renders alternative explanations implausible only if it is still effective after the randomization. This effectiveness is lost, e.g., if subjects who have been assigned to different groups interchange information about the experiment. Further, subjects in a group might be dissatisfied with their experimental condition and find better conditions, which are similar to the conditions for other groups, outside the experiment. Finally, a randomization may not work because subjects, in contrast to a former promise, refuse to participate in an experiment after they have learned about the condition they were assigned to by the randomization procedure.

Answers to Chapter 2

2.1 Instead of the term "dependent variable" we can use the term "effect variable", and instead of "independent variable" we can use "causal variable".

2.2 If one wants to check whether a psychotropic drug reduces anxiety, the dependent variable "anxiety" obviously is a construct or a latent variable, respectively, which permits no direct measurement. One tries therefore to find a suited operationalization for the latent variable "anxiety", which can be recorded directly, e.g., the score of an anxiety questionnaire or the heart-beat rate.

2.3 a. One control condition should be in any case a level of the independent variable for which no effect on the dependent variable is expected.

b. It should be tried to fix an extreme level by requiring that just at this level no harmful effect on the subjects is suspected but for all more extreme levels.

c. It should be tried to find a smallest level where just still an effect can be expected but for no smaller level.

d. If a natural ordering of the levels is possible, e.g., for doses of a drug, the distance between two succeeding levels should be so large that different effects can be expected.

e. The levels should be selected in such a way that the whole effective range of the independent variable is covered without too large gaps.

f. If it is known in case of ordered levels where small changes of the levels are accompanied by large changes of the effect, the distances between the levels should be chosen smaller in such regions.

g. Without foreknowledge it is advisable to keep the distances between the levels constant as far as this is possible. Here, "constancy", e.g., for drugs, may also mean "logarithmic constancy" if it can be assumed that only a linear increase of the effects is observed if the dose is raised to a higher power. In such a case, it might be wise to choose each dose twice as high as the preceding dose.

2.4 If the experimenter cannot choose the levels of an independent variable in an arbitrary fashion but must select them from a set of given levels, it is not possible to assign subjects randomly to these levels. If, e.g., it is asked whether the causal variable "smoking" has the effect "lung cancer", the populations of "smokers" and "non-smokers" pre-exist. If it would be really possible to draw from each of the two populations a random sample, it could be concluded, whether lung cancer occurs with a higher probability in smokers than in non-smokers. Then the following causal conclusion is possible: The probability of lung cancer is increased for smokers. However, the following conclusion would not be permitted: Smoking increases the probability of lung cancer. This latter conclusion would be unfounded because it may be that not smoking causes lung cancer but something else (e.g., a gene) which at the same time causes lung cancer and drug dependence.

Because it is not possible in practice to draw true random samples from the populations of smokers and non-smokers, always selection effects must be expected, i.e. it can never be ruled out that the samples of smokers and non-smokers do not differ solely with respect to smoking behavior but that they are also different with respect to other variables in a systematic way and that one of these other variables causes lung cancer. It might be possible, e.g., that the subjects in the considered sample of smokers have inhaled in their childhood carcinogenic chemicals in contrast to the subjects of the considered sample of non-smokers. Such difficult to disprove alternative explanations can be ruled out only then, at least in the statistical sense, if subjects are randomly assigned to the conditions smoking and non-smoking and are not selected from pre-existent populations.

2.5 A high positive correlative relation means that high values of a variable A coincide with high values of a variable B and that likewise low values of a variable A coincide with low values of a variable B. Correspondingly, a high negative correlative relation means that high values of a variable A coincide with low values of a variable B and low values of a variable A with high values of a variable B.

One possibility for the occurrence of such correlative relations are causal relations. If, e.g., smoking really causes lung cancer, we should have a high risk of cancer for smokers and a low risk for non-smokers, i.e. altogether a high positive correlation between smoking behavior and the incidence of lung cancer. However, such a high correlation can be found also for other reasons: a certain gene might affect that persons at the same time are addicted to smoking and get lung cancer though smoking does not increase the risk for lung cancer. Another possibility is that lung cancer is caused by inhaling certain substances in childhood and that corresponding subjects are overrepresented in the sample of smokers under consideration. Again, a positive correlative relation between smoking behavior and the incidence of lung cancer would be observed. Thus, correlative studies can never be used for proving the existence of causal relationships.

2.6 If a cause is identified that is responsible for an effect this does not mean that this cause is directly responsible for the effect. It is possible that the cause influences an intervening variable which in this sense is a dependent variable. If this intervening variable is responsible for the effect which was observed at the beginning, the intervening variable is an independent variable with respect to this effect. If in this way one or more intervening variables are active between cause and effect, we have a causal chain. An example for such a causal chain is the influence of the price for cat-food on fruit-crop in Section 2.3.

As a rule, intervening variables or causal chains are only considered if a detected causal relation seems puzzling. However, this does not mean that there do not exist intervening variables also in cases where the observed causal relations seem to be plausible to us.

2.7 To operationalize the construct "nervousness" for a subject we might think of the following variables:

a. The chair of the subject is equipped with sensors which cannot be perceived by the subject and which record changes of bearing.

b. The subject is asked to track with a pencil a curve and the number and extent of deviations from the curve are recorded.

c. A text has to be read aloud by the subject and the number of slips of the tongue is counted.

2.8 Ockham's razor says that one should always choose the simplest explanation if more than one explanation is available for a phenomenon. Because of the obvious pragmatism of this principle it is very tempting to apply it whenever several explanations are possible. However, in practice it is often not clear what the simplest explanation might be. There may be people to whom explanations based on drives, psychic forces, ghosts etc. might appear as much more simple than any scientific approach. Further, one cannot rule out that it seems that a given phenomenon can be explained by a simple causal relationship though it is in reality the result of a complicated interplay of several causes.

2.9 Here, a correlative study was performed, i.e. it was only observed whether two dependent variables have high or low values at the same time. Due to this design it is not possible to conclude whether one of the two variables is a cause for the other one. By no means it can be ruled out that there exists an independent variable which is not considered here which is responsible for the observed correlations. If this causal variable would be kept constant the apparent relation between the two studied variables would disappear. Therefore, such correlational studies are of no use for studying the actually existing relations between two constructs.

If one would like to study, e.g., whether conservatism has an influence on the acceptance of incongruity humor, a sample of subjects might be randomly split up into two subsamples. Each of the two samples should be isolated for a longer period from the outside and undergo an indoctrination where one group should learn conservative, the other group non-conservative conceptions of values. After this both groups had to answer the humor test. A difference in the test scores of both groups could be an indication that a conservative or non-conservative, respectively, view of life yields a different judgement of the funniness of incongruity humor. Here, it is assumed that conservative and non-conservative views may be produced by indoctrination.

Against the indoctrination of subjects proposed for the design above, scruples may be advanced for ethical and possibly also for legal reasons. These might possibly be ignored, however, in case that the participation in the study is completely voluntary. After all, there seems to be no country on earth where at the present moment people are not indoctrinated in a partial way by governmental institutions, by media, or by ideologically committed groups, without that these people have really taken a free decision.

Answers to Chapter 3

3.1 We assume that a causal variable influences an intervening variable which in return leads to the observed effect. Furthermore, we assume that the causal variable causes the effect also directly, i.e. that the causal relation is observed even if the intervening variable has been eliminated. In this case the intervening variable is not a confounding variable according to our definition.

A little child might start to cry, e.g., if a car's door is slammed or if a dog barks. Always when a car's door is slammed the dog barks. Thus, we have a direct causal relation (a car's door is slammed → the child cries) and a causal chain (a car's door is slammed → the dog barks → the child cries). Obviously, the direct causal relation is maintained even if the dog is eliminated from the neighborhood of the child and the causal chain is interrupted. Thus, the dog is not a confounding variable. This would be the case, however, if only the dog but not the child could hear the noise produced by slamming a car's door. Then only the causal chain would exist but the direct causal relation would not.

3.2 If it is not possible to prove the existence of causal relations by means of a certain study, i.e. if this study has no internal validity, it cannot contribute to our knowledge about relations in the real world because, e.g., observed correlative relations permit many possible interpretations. Such a study is worthless even if it is performed for many populations of subjects and for many different situations in order to increase the external validity. If a study does not increase knowledge a generalization of this ignorance is meaningless.

A typical example for this kind of proceeding is the large number of studies in Personality Psychology, where the "big five factors" were detected in samples of subjects from many populations. These are pretended "personality factors" which are derived by so-called factor analyses from correlations of variables measured by questionnaires. If one assumes a certain similarity of the structure of questionnaire data which were obtained in different populations, the application of the same mathematical procedure will produce similar results in most cases. However, this is not a proof for the actual existence of the described personality factors, which may be simple artifacts.

3.3 A null result is present if no causal or correlative relation can be proved. In most cases such a null result is present because a statistical test yielded a result which was not significant, i.e. where a null hypothesis could not be rejected. A null result, in principle, cannot be interpreted because one will never be able to rule out that a relation might be found in a future study, possibly with a larger sample size. No conclusions can be drawn from such null results with respect to any theories. As for the planning of future studies one must not assume either that effects whose existence could not be proved, do not exist.

3.4 If a low statistical power is suspected, the following strategies may be used to prove the existence of relations:

a. The sample size can be increased.

b. The reliability of the used dependent variables can be increased or other, more reliable dependent variables can be used.

c. The results of several independent studies can be combined, e.g., with the method described in Section 9.4.

3.5 If the subjects' behavior at later measurements is influenced by the effect of the measurement at a first point of time, this is called sensitization. Thus, a first taking of a blood sample will probably influence the behavior of subjects if blood samples are taken at later points of time. In such cases, we have reactive or obtrusive measurements. However, if subjects are being observed though they do not know this, one cannot assume that a former observation influences the behavior of subjects at later observations. Here, we have non-reactive or unobtrusive measurements.

3.6 a. The existence of an effect of the method cannot be proved without the use of a control group. The original sample should be randomly split up into two subsamples, where the new method is used in one subsample, and a conventional learning method in the other subsample. If no control group is being used any observed changes from pretest to posttest might have been caused, e.g., by history (cf. Section 3.2.1) or maturation (cf. Section 3.2.2). Without a control group a selection effect (cf. Section 3.2.6) cannot be ruled out either, as the researcher obviously expects a measurable "success" of the new method only for the children with lower scores. Finally, one cannot rule out that the new method might even produce negative changes for children with average or high scores.

b. One should dispense with the pretest not only because a regression effect might occur, but also because misinterpretations due to testing (cf. Section 3.2.3) and instrumentation (cf. Section 3.2.4) are possible.

3.7 A measuring instrument which serves as a means to record the extent of dementia might consist of a list of six words denoting objects from daily life. During the test the list is once read slowly to the subjects and they are then asked to recall as many of the words as possible from their memory. A sample of dements and a control sample of depressives is used. After a pretest a memory training is performed with the dements. After the training both groups are tested anew. Because the pretest shows a distinct superiority of the depressives while in the posttest both groups exhibit comparable outcomes the training is believed to have helped the dements to achieve a normal memory performance.

This conclusion should be regarded with some scepticism because one cannot rule out that a ceiling effect is present, which worked in the pretest only for the depressives but in the posttest for both samples. If a list with 20 words had been used we might have found an average outcome of 12 remembered words for the depressives in the pretest and posttest, while the dements might have remembered an average of two words in the pretest and of six words in the posttest. Due to instrumentation one would not have revealed that the performance of these samples from two different populations cannot be compared.

In this example it is not allowed to conclude from the difference of the performances of the dements before and after the training that this difference is an effect of training. Because no suitable control group is used one cannot rule out that merely the increased attention shown to the dements has improved the performance. A causal conclusion would have been feasible if the original sample of dements had been randomly split up into an experimental and control group.

3.8 Causal conclusions with respect to the construct "pain" can be derived only if this construct has been defined in a sufficiently precise and restricted manner. First, one has to distinguish "emotional pain" and "physical pain" where only the latter can be operationalized with sufficient precision in an experimental situation. E.g., quantitatively graded levels of "pressure pain" can be produced by counting the number of turns of a thumbscrew, which is fastened to a precisely defined location of the right thumb of a subject (if the subject is right-handed). Quantitatively graded levels of "burning pain" can be produced by measuring the duration (in seconds) for which a metal rod which is heated to a defined temperature is pressed on a precisely defined location of the back of the right hand of a (right-handed) subject. The subjective sensation of pain might be measured by a rating scale, where the subjects have to name a number between zero (corresponding to no pain) and 100 (corresponding to intolerable pain).

If one wants to know whether listening to classical music has a soothing effect, a sample of subjects might be randomly split up into two subsamples of which one listens to a certain piece of classical music while the other does not get any acoustic stimuli. Pain is inflicted to the subjects in both samples in the same way and to the same physical extent. Each subject rates the extent of subjective pain. If both samples differ only in listening or not listening to the piece of classical music, a comparison of the subjective pain scores for the two groups may permit a statement of the kind "listening to the piece *X* of classical music reduces the subjective sensation of pain of modality *Y* and of the physical level *Z*".

3.9 Noise might be presented via
a. headphones,
b. loud-speakers.

As noise one might use
a. traffic noise,
b. aircraft noise,
c. industrial noise,
d. natural noise, e.g., animals in a zoo or on a farm,
e. artificial noise, e.g., white noise,
f. instrumental light music,
g. vocal light music,
h. instrumental classical music,
i. spoken text, e.g., a newscast.

Memory performance can be measured by means of
a. reproduction,
b. recognition,
c. relearning.

One could make subjects learn the following material
a. nonsense syllables,
b. syllables with meaning,
c. texts with meaning,
d. abstract symbols, e.g., artificial letters,
e. pictures depicting concrete objects.

One might use
a. men, women,
b. children, adolescents, grown-ups, old people,
c. healthy people, dements, schizophrenics, depressives
as subjects,

and
a. early morning,
b. noon,
c. afternoon,
d. evening,
e. night
as experimental times.

The experiment might take place in
a. plain laboratory rooms without stimuli,
b. rooms with distracting stimuli, e.g., an interesting interior or with windows opening to a thoroughfare.

3.10 If one wants to avoid that the responses of subjects are influenced by their desire to appear as normal, healthy, able, and intelligent as possible, to the experimenter or to the subject evaluating the data, various situations can be distinguished:

a. Alterations of the outcomes due to social desirability responding are no threat to construct validity if the subjects in the study cannot influence the measured dependent variable. This is the case, e.g., if the measurements are recorded without knowledge of the subjects or if the measurements, e.g., blood pressure, cannot be influenced deliberately.

b. "Social desirability responding" will have little effect in the case where different levels of the dependent variable have no relation with the image, the subjects desire to present to other people. The question whether a subject can work better in the morning or in the afternoon might be an example. However, this very question will be answered free of social desirability only if it is asked in the context of a general interview or a scientific study. If this question is posed in the context of an employment or during negotiations concerning employment, social desirability responding cannot be ruled out.

c. Very often a subject believes, justified or unjustified, that certain levels of a recorded variable may have the consequence that the image that other people have of the subject has changed into an unwanted direction. In this case an alteration of the outcome by social desirability responding has to be taken into account. To avoid corresponding response biases sometimes the **dark room effect**, as we will call it, can be utilized. This means that a group of subjects in a completely dark room may exhibit a behavior different from the behavior of the same group in a lighted room. The reason for the different behavior in the two situations might be found in the fact that subjects in a dark room do not expect the same actual or imagined social sanctions for a certain behavior as they would expect in a lighted room.

If we transfer this principle to an interview situation we might form groups which are homogeneous with respect to age, gender, and other external characteristics which may be important for the social status. Such a group, say of 20 subjects, finds 20 copies of a questionnaire, 20 envelopes, and 20 ball pens at the interview location. The subjects distribute these objects by a random procedure in the absence of an experimenter, assigning to each subject a questionnaire, an envelope, and a ball pen. Each subject then fills out the questionnaire, shielded from the others, and puts it into the envelope. Finally, all the subjects together perform a random shuffling of the envelopes. As the subjects know of this procedure before participating in the study, they are sure that the answers cannot be assigned to the corresponding subjects. Thus, not only the fear of possible sanctions would not have any basis, but also every incentive for social desirability responding because it would not be for the benefit of the subject giving the answers. This proceeding would fail, however, with paranoide (or very intelligent) subjects who suspect that the other 19 subjects are confidants of the experimenter.

d. It is particularly difficult to control social desirability responding if an interaction with the treatment factor cannot be ruled out, e.g., if a higher extent of bias in the outcomes has to be expected under the treatment than under the control condition.

e. In this context, selection effects cannot be ruled out either. Thus, men might perceive the same subjective pain for a certain physical stimulus as women but, due to their gender role, only admit a lower effect when pain is being measured.

f. In particular, in case of achievement tests, e.g. intelligence tests, the effect of social desirability responding is taken advantage of to motivate the subjects. They try to achieve optimum results, assuming that a high ability goes hand in hand with a high social acceptance. This motivation technique fails, if a high social acceptance comes along with a low utility in other respects. E.g. somebody, who wishes to get money from an insurance company, will rather try to present low achievement scores.

3.11 The most obvious proceeding to prove the existence of a Rosenthal effect consists in the introduction of expectancy control groups as discussed in more detail in Section 4.10.5. In the simplest form a sample is randomly split up into three subsamples, where in addition to a treatment and a control group an additional control group is used where implicitly an expected effect is suggested to the experimenter who is in contact with the subjects. One could declare, e.g., that in this very group several subjects are present which had very high scores in a preceding intelligence test. Or, in an animal study one could inform the experimenter that in the corresponding subsample some animals from a litter are present, of which the parent animals achieved very good results in a former experiment. If a distinct difference between the outcomes of the two control groups was found though the only difference between the two groups is the presence or absence of the (false) additional information for the experimenter, the difference of the outcomes can be ascribed to a Rosenthal effect.

3.12 The aim of a single-blind study with subjects is to guarantee that the expectancies of the participating subjects are not connected in a systematic way with the levels (or level combinations) of the independent variables. Otherwise, one cannot decide if an effect is due to the conditions used by the experimenter, to the expectancies of the subjects or to both.

In case of a single-blind study with animals, the animals cannot distinguish the different levels (or level combinations) of the independent variables. By this method one wants to achieve that effects are only due to relevant aspects and not to irrelevant aspects of the levels of the independent variables.

In a double-blind study with human beings, neither the respective subject nor the experimenter does know which levels of the independent variables are effective. By this method one intends that neither the expectancies of the participating subjects nor the expectancies of the experimenter have a systematic influence on the outcome.

In a double-blind study with animals, the experimenter does not know which levels of the independent variables are present for the single animal and the animals cannot distinguish the different experimental conditions. By this method one wants to achieve that expectancies of the experimenter do not influence the outcome and that only relevant aspects of the experimental conditions can have caused detected effects.

In a triple-blind study with subjects, neither the respective subject nor the experimenter does know which levels of the independent variables are effective. Further, the subject who evaluates the study only knows which groups are to be compared but not which experimental conditions were assigned to the different groups and in which way. By this method one wants to achieve that neither the expectancies of the participating subjects, nor the expectancies of the experimenter nor the expectancies of the evaluating subject can influence the outcomes in a systematic way.

In a triple-blind study with animals, the animals cannot distinguish the different experimental conditions. The experimenter does not know which levels of the independent variables are present for the single animal, and the subject who evaluates the study only knows, which groups have to be compared but does not know which experimental conditions were assigned to the different groups in which way. By this method one tries to avoid that expectancies of the experimenter influence the outcome, that expectancies of the evaluating subject have an influence on the selection of evaluation procedures and the interpretation of the results, and that, finally, effects are not only due to relevant differences of the experimental conditions.

3.13 A single-blind study should be used, where the experimenter does not know the experimental condition to which the respective animal is exposed. This could be realized, e.g., in studies with "counting" animals as the Clever Hans in the following way:

Before the experiment, the experimenter records the questions, which the animal will be asked, in a standardized form on tape. The numbers used in the questions are to be selected randomly from the set of all admissible numbers. During the experiment the experimenter is separated from the animal by a transparent but sound-absorbing pane, i.e. the animal sees but does not hear the experimenter. During the experiment the experimenter asks the questions. However, the animal does not hear these questions but instead via a loudspeaker a random sequence of the questions recorded in advance. As a kind of control, in some cases the two questions should coincide. If one found out that the animal gives correct answers to the questions transmitted via loud-speaker only in the control condition and in the other cases answers to the questions of the experimenter which the animal cannot hear, this would indicate that the animal responds only to optical stimuli which are given intentionally or not by the experimenter. This means that the animal is not able to understand the questions and, therefore, cannot answer them correctly.

A counter-argument against this kind of reasoning might be that animals understand the human language and, thereby, the questions, not acoustically but, similar to deaf subjects, by lip-reading. In case of such an argumentation one might test whether the animal can still answer the questions if it sees during the time when the question is asked only the head of the experimenter. This head is fixed in such a way that it cannot be moved and after the question is asked, the head is replaced by a wax version of it without the animal being able to perceive this change. Further, the experimenter should have been trained in advance to ask all questions in a strictly standardized verbal form, where only the respective numbers are altered. If the animal is still able to produce the correct answers, an alternative explanation might be that the experimenter transmits the correct solution while asking the question by means of mimic or lip movements. This source of bias is difficult to eliminate. However, when reading the report of Oskar Pfungst (1907/1977) about the abilities of Clever Hans one learns that, obviously, nonverbal information about the solutions was always given after the question had been asked, i.e. with our experimental proceedings the tasks would have been too difficult for Clever Hans.

3.14 A sample of subjects is randomly split up into two subsamples one of which gets a drug, the other a placebo. The object is to prove the existence of potential side-effects of the drug. Before drug or placebo is being applied, the subjects get a questionnaire in which they are asked for 14 symptoms:

Check off, which symptoms you have perceived during the preceding 45 minutes:

○ itching of the skin
○ itching of the eyes
○ hunger
○ thirst
○ micturition
○ dizziness
○ respiratory disturbances
○ cardiac rhythm disturbances
○ uneasiness

○ restlessness
○ depressive thoughts
○ increased salivation
○ short-time absent-mindedness
○ day-dreams

One hour after they have received drug or placebo the subjects get the questionnaire again. It would be no surprise if both groups reported none or only few symptoms in the pretest, but several symptoms in the posttest. The subjects were sensitized by the pretest to which symptoms they should pay attention and, therefore, they will "detect" at least some of these symptoms under both conditions, in particular, in view of the relative long duration of one hour between the two measurements. If only a posttest had been used, the chance would have been higher that only those symptoms were reported which actually were present. Then, a difference between the two groups might be observed.

However, in this example one cannot rule out completely that the subjects might have reported about symptoms which did not appear in reality by the mere awareness of being in an experimental situation, also in case of a posttest without a preceding pretest. This again might affect an assimilation of the outcomes for both groups.

Answers to Chapter 4

4.1 The two factors "number of syllables" and "meaning of syllables", each with two levels, yield the design in Figure 4.23. Note that only the two comparisons in each row or in each column, respectively, i.e. altogether four comparisons, make sense, while the two "diagonal comparisons", as discussed for Figure 4.4, yield results which are difficult to interpret.

	Meaning of Syllables	
Number	Nonsense	With Meaning
7	7/Nonsense	7/With Meaning
14	14/Nonsense	14/With Meaning

Figure 4.23: Level combinations of the factors "number of syllables" and "meaning of syllables"

In contrast to the design in Figure 4.3 the two factors in the design depicted in Figure 4.23 can be effective at the same time. With the design in Figure 4.23 one can investigate which effects can be expected if number and meaning of the syllables to learn are being considered at the same time. Compared to this, a design as depicted in Figure 4.3 or the designs depicted in Figure 4.5 and 4.6 can be used if the after-effects are of interest which can be produced by an alteration of the number of nonsense syllables or of the number of syllables with meaning, respectively.

4.2 Four opaque synthetic globes of the same size are produced in which lead bullets are contained such that the globes have a mass of 80 g, 160 g, 640 g, and 1280 g. We consider the two factors "small mass" with the two levels "80 g" and "160 g" and "large mass" with the two levels "640 g" and "1280 g". An original sample of right-handers is randomly split up into four subsamples which are assigned to the four level combinations (80 g, 640 g), (80 g, 1280 g), (160 g, 640 g), and (160 g, 1280 g). Each subject has to take the two globes, one after another, which are assigned to him or her into his or her left hand and rate the mass (in gram) of each single globe, i.e. we consider here two dependent variables.

It is obvious that the two factors cannot be realized at the same time. If always the smaller and than the larger mass is rated one can assume that more precise ratings result than in the opposite case. For, if always first the larger and then the smaller mass is rated one cannot rule out that the smaller mass is underestimated. A reason for this effect of order might be that the reference point for the succeeding subjective mass rating is considerably more shifted by a large mass than by a small mass. This consideration could have the consequence that the small mass is always to be rated first followed by the large mass.

	Phase 2	
Phase 1	80 g	1280 g
80 g	80 g / 80 g	80 g / 1280 g
1280 g	1280 g / 80 g	1280 g / 1280 g

Figure 4.24: Design for the investigation of the supposed effect of order for the factors "small mass" and "large mass"

An experimental verification of the supposed effect of a mass rating on the reference point could be performed with the design in Figure 4.24 where only the smallest mass (80 g) and the largest mass (1280 g) are being considered. The supposed effect would predict that the rating score for the mass 80 g in phase 2 is smaller for group (1280 g/ 80g) than for group (80 g/80 g) and that the rating score for the mass 1280 g is larger for group (80 g/1280 g) than for group (1280 g/1280 g). However, for the first comparison a larger bias has to be assumed.

4.3 If, in Figure 4.5, only the outcomes after phase 1 are of interest, no problems will arise. It is even possible to pool two accordant groups with four subjects each to one group with eight subjects, if the conditions in phase 1 are the same for this group. Thus, a design with two factors results, with the two levels (7, 14) or (nonsense, with meaning), respectively. However, an interpretation of the outcomes after phase 2 is only possible for groups where we have the same condition in phase 1. Otherwise, it will not be known whether effects which are found in phase 2 are due to differences of the conditions in phase 2 (as hoped) or to differences of the conditions in phase 1 or to differences of the conditions in both phases. As each condition occurs only twice in phase 1 in Figure 4.5, at a time only one of the four theoretically possible comparisons for the four conditions in phase 2 can be performed without problems of interpretation. All these 4 × 4 = 16 valid comparisons can be performed by using the design depicted in Figure 4.6.

4.4 One has to consider a total of 4 × 16 = 64 level combinations since each of the 16 combinations in Figure 4.6 has to be combined with the 4 combinations (7, nonsense), (14, nonsense), (7, with meaning), and (14, with meaning) in phase 3.

4.5 According to the proposal of Riecken et al. (1974, p. 175), subjects should be randomly assigned to different conditions only if they have been informed about the possible alternatives and have consented to participate in the study to whatever condition they will be assigned to. If one assumes that the different conditions correspond to different therapies, which moreover can be discriminated by the subjects, i.e. if no blinding is possible, a conventional therapy might appear less risky to most subjects. Cautious patients will therefore be inclined to refuse to participate in the study if they are assigned to a new therapy. Compared to this, more courageous patients will remain in the study, even if they are assigned to a new therapy. This yields a selection of the form that there will be more courageous patients in the sample with the new therapy than in the case of a true random assignment.

Such a bias in the composition of the samples can result in a selection effect in more than one respect. On the one hand the courageous patients may be persons who are generally inclined to take greater risks and who will not be as cautious during therapy as it would be necessary. But these patients might also be more courageous because they set all hope on the new therapy due to their past experience and the seriousness of their illness. In both cases, the percentage of patients with a higher risk would be larger in the sample with the new therapy than in the sample with the conventional therapy, because randomization has not worked.

4.6 A new and a conventional therapy for treating breast cancer are to be compared with each other. According to the proposal of Riecken et al. (1974, p. 175) the patients are informed about the risks and chances of both therapies and are consequently asked whether they will participate in the study irrespective of which of the two therapies they are assigned to. After giving their consent, the patients are randomly assigned to the two therapies. After the assignment, some of the patients who were assigned to the new therapy and possibly also some of the patients who were assigned to the conventional therapy refuse to take further part in the study. As discussed above, the result of a comparison of the two therapies does not permit causal conclusions under these circumstances, as those patients remaining in the two samples may differ not only with respect to their therapy but also, e.g., with respect to the risk for a possible failure of a therapy.

Compared to this, according to the proposal of Zelen (1979) the random assignment of patients to two samples is performed though the patients have not been informed about possible alternatives and have not been asked for their consent. The patients of the one sample get the conventional therapy, i.e. they are treated in the same way as patients who did not participate in the study. The patients of the other sample are asked whether they consent to be treated with a new therapy. If this is the case, they are correspondingly treated, otherwise they get the conventional therapy. All patients of this sample are, of course, informed about the risks and chances of both therapies before they are to take a decision. At the end the outcomes of all patients in the second sample are compared with the outcomes of all patients in the first sample. Here, a causal conclusion is possible because systematic selection effects can be ruled out due to the performed randomization. The only drawback might be that a difference between the effects of both therapies cannot be detected because too few patients have consented to be treated with the new therapy.

4.7 In the beginning, there are $u = 3$ cards in the box, on which is written an A, and $u = 3$ cards on which is written a B, i.e. there are altogether 6 cards in the box. Therefore, for patient 1 the probability to be treated with therapy A is given by $(3 / 6) = .5$ and to be treated with therapy B is also given by $(3 / 6) = .5$. Because patient 1 corresponds to case 1, an A-card is drawn, the patient is treated with therapy A, and his or her treatment is successful. Therefore, $v = 4$ A-cards and $w = 2$ B-cards are additionally put into the box. This now contains 7 A-cards and 5 B-cards. Thus, for patient 2 the probability to be treated with therapy A is given by $(7 / 12) = .583$ and to be treated with therapy B by $(5 / 12) = .417$. Because patient 2 corresponds to case 4, a B-card is drawn, therapy B is applied, and this therapy has no success. Therefore, again $v = 4$ A-cards and $w = 2$ B-cards are additionally put into the box, which now contains 18 cards, namely 11 *A*-cards and 7 *B*-cards.

Thus, the probability for an *A*-card is equal to (11 / 18) = .611 and for a *B*-card (7 / 18) = .389 for patient 3. Again the patient corresponds to case 4, i.e. again v = 4 *A*-cards and w = 2 *B*-cards are added. This yields 24 cards altogether, namely 15 *A*-cards and 9 *B*-cards. Therefore, for patient 4 the probability for an *A*-card equals (15 / 24) = .625 and for a *B*-card (9 / 24) = .375. Because patient 4 corresponds to case 2, for this patient an *A*-card is drawn, therapy *A* is applied, but this therapy is not successful. Therefore, now w = 2 *A*-cards and v = 4 *B*-cards are added, which yields altogether 17 *A*-cards and 13 *B*-cards. For patient 5 the probability for therapy *A* would equal (17 / 30) = .567 and for therapy *B* (13 / 30) = .433.

4.8 A covering stimulus may be that strong that it covers not only the disturbance stimuli (as desired) but also those stimuli, which correspond to some or even to all levels of an independent variable.

E.g., the influence of music styles on the construct "attention" is to be studied. For this, the independent variable "music style" is considered with the two levels "classical music" operationalized by a piano sonata by Beethoven and "light music" operationalized by a song of the Rolling Stones. A sample of subjects is randomly split up into two subsamples and each subsample is assigned to one of the two levels. Symbols move over the screen of a personal computer and each time when one of two given symbols appears, the subject has to press a key during the time interval where the symbol is perceptible. The number of omitted symbols and the number of falsely identified symbols serve as dependent variables.

In order to cover acoustic disturbance stimuli as, e.g., noise caused by movements of the experimenter, opening of the door, movements of the subject's chair, starting of a ventilator etc., a tone generator produces a constant noise. If this noise is too weak, the background noise may influence the attention of the subjects in an uncontrolled way, such that no causal relation can be detected between music style and attention. However, if the covering noise is too strong, the two levels of the factor "music style" can no longer be perceived as different, such that, for this reason, a causal relation cannot be detected. Finally, it may be that the covering stimulus has a strength where "classical music" in contrast to "light music" can no longer be perceived. Again, a causal conclusion is impossible because the effects of the two styles of music cannot be compared with respect to their effect on attention due to the covering of the "classical music".

4.9 a. If subjects participate in an experiment, they should be given instructions in a very objective way, with the proceeding being the same for all subjects. If different instructions are used for different levels or level combinations of independent variables, i.e. also for different subsamples, one can no longer rule out that observed effects are not caused by the independent variables but are due to the differences in the instructions. This holds also for very small formal differences of the texts. The subjects should get their instructions without a personal contact with an experimenter, if possible in written form. If all subjects get exactly the same instructions, which is the best thing to do, these can also be given in acoustic form via a tape.

b. Of course, one cannot give instructions to animals in verbal form. Here, a shaping procedure must be used. For this, the natural behavior set of the animals is used by reinforcing step by step those behaviors which are important for the experiment. These may be, e.g., the alternate pressing of two levers for rats or pecking on lighted disks for pigeons. If the required behavior is too much unlike the natural behavior of the animals, sometimes an additional priming is used, i.e. the animals are brought passively to the desired activity, sometimes by the hand of the experimenter, to experience then the consequence, i.e. a reinforcement.

While shaping is a well-known technique of animal training, it is sometimes overlooked that little children who are still not able to understand verbal instructions, acquire desired behavior, e.g., when learning to speak, by shaping. Also, in case of certain psychotherapies, e.g., therapies used for treating stutterers, shaping techniques are employed.

4.10 If it is not possible to prove the existence of causal relations by a study, one possible reason may be that the participating subjects are so heterogeneous that existing differences, e.g., between an experimental and a control condition, cannot be detected. One way to solve this problem is to select a single sample of subjects, who are all very similar with respect to certain block variables, which the researcher judges to be important. Then, this sample is randomly split up and assigned to the different experimental conditions.

This proceeding corresponding to a global homogenization has two disadvantages: First, it will be difficult, in general, to find a sample which is sufficiently large on the one hand and sufficiently homogeneous on the other hand. Second, any detected causal relations can only be generalized to a very restricted population.

Both disadvantages can be avoided in case of matching or blocking. Here, the original sample of possibly very heterogeneous subjects is split up into subsamples, which are as homogeneous as possible with respect to one or more matching or block variables. This homogeneity of the subjects of a subsample should also hold for the dependent variable, which still has to be measured. While the subjects of a subsample should be very similar to each other, the subjects of different subsamples can differ considerably. Each subsample is randomly split up and the resulting parts are randomly assigned to the different experimental conditions. Within each subsample it might be easier to detect causal relations because of the homogeneity of subjects within the subsample, than without matching or blocking. Because the subjects of different subsamples may differ considerably, only causal relations, which were found for several subsamples, can be generalized to a heterogeneous population.

Beside the obvious advantages of matching or blocking (facilitated detection of causal relations, improved generalizability of the results) this technique has also several disadvantages:

a. There always exist infinitely many potential matching or block variables, which might contribute to the heterogeneity of subjects. Many of these variables are unknown and even known block variables can often only be measured with high expenditure and doubtful reliability.

b. If unsuited matching or block variables are being used, i.e. if the subjects within the subsamples are homogeneous with respect to these variables but not with respect to the dependent variable which is to be measured, matching or blocking does not only not facilitate the detection of causal relations but even renders it more difficult. This impediment has its reason in the fact that appropriate statistical procedures have to take into account that only a restricted randomization has been performed within the subsamples and no global randomization. For such procedures the probability to detect existing effects decreases if only an insufficient homogenization of the samples is achieved with respect to the dependent variable before the independent variable becomes effective. The outcome may be even considerably worse than in the case of a global randomization where the original sample is randomly split up and assigned to the experimental conditions without forming homogeneous subsamples in advance.

   It would not be advisable in a case like that to simply "forget" that pairs or blocks were formed and to use a statistical procedure which assumes a global randomization. Then, it may happen that the existence of not existing causal relations is "proved" or that existing causal relations are not being detected.

c. When forming homogeneous pairs or blocks, it may happen that one does not find a partner which is sufficiently similar for certain subjects. If these subjects without partners are left out, selection effects can occur. These have the effect that possibly causal relations, which were detected for subpopulations, are generalized to the total population without justification.

d. If matching or block variables cannot be measured without errors this can have the effect, just as a too low correlation with the dependent variable, that the derived homogeneity in the subsamples is not achieved with respect to the dependent variable. This again renders the detection of causal relations difficult.

e. In case of more than one matching or block variable it is often impossible to find suited subjects for all possible level combinations of these variables, such that selection effects may arise as in argument c above.

f. If matching or blocking is used, an overmatching can take place where a matching or block variable is a variable within the causal chain between the independent and the dependent variable. Local constancy of this variable prevents the detection of existing causal relations.

g. Because matching or block variables are measured before measuring the dependent variable, sensitization effects cannot be ruled out which complicates the detection of causal relations.

4.11 A randomization after matching meant in this case that 100 pairs of patients had been formed before an assignment to the clinics, where the patients within a pair had not differed with respect to the level of the matching variable "chemotherapy" and with respect to levels of other possible matching variables. This means that both patients of a pair would have got chemotherapy or none of the two. Within each pair one patient would have been randomly assigned to clinic A and therefore the other one to clinic B. If the perceived quality of life of the patients is essentially influenced by the use or non-use of chemotherapy, most probably a difference between the two clinics cannot be found. Thus, as it could be expected, we would have got the same null result as in the case of a randomization before matching.

The described matching procedure is not realistic, because it will not be possible for ethic as well as for legal reasons to prescribe to the clinics which patients will receive a chemotherapy and which will not. This is particularly true, if this is a decision which is not based on a medical diagnosis but on the result of the toss of a coin. However, if the matching variable "chemotherapy" is not considered in our example, a superiority of clinic A is found as it was the case without matching at all. If matching variables such as age, gender or seriousness of illness are considered and if one assumes that the quality of life is different for different level combinations of these variables the chance to detect a difference between the clinics is increased, which may be essentially caused by the frequency of the use of chemotherapy.

To avoid any misunderstanding: the proceeding of researcher Y, who performed a matching after a randomization, cannot yield outcomes that permit causal interpretations because one can never be sure that the patients do not only differ with respect to the clinic but also with respect to many other unknown variables. Such differences can, e.g., have the effect that an existing clinic effect is not detected, which is a possible alternative explanation for the outcome reported in the text. For similar reasons, however, differences between the samples might be found though no clinic effect exists.

4.12 a. One is not told, whether a randomization was used, i.e. whether the partitioning of the sample of 84 subjects into 6 subsamples each with 14 subjects was done randomly or not. Without a randomization, observed group differences may be solely due to selection because the subsamples can differ in a systematic way even before the experimental conditions have been introduced.

Furthermore, obviously no baseline phase, to measure the ability of the subjects, has been introduced before the experimental phase. With such a baseline it would have been possible to match subjects of the experimental group and of the yoked control group before randomization as it is common practice. This might have facilitated the task of detecting group differences.

b. We make the assumption that with an increasing number of aversive tones the dependent variable "cardiovascular reactivity" is increased for "sensitive" subjects while no effect is found for "insensitive" subjects. Here, "insensitiveness" with respect to aversive tones can be caused, e.g., by hardness of hearing. In order to simplify the discussion and also in order to avoid too many cases being considered, we assume that the ability of a subject in the experimental group does not differ from that of its partner in the yoked control group. To achieve this, a matching might have been advisable.

Now we can discriminate four situations:

1. Both subjects of a pair are "sensitive". Then a comparable high increase of the dependent variable is observed for both subjects.

2. Both subjects of a pair are "insensitive". Then no increase of the dependent variable will be observed for any of the subjects.

3. The subject in the experimental group is "insensitive", the corresponding subject in the yoked control group is "sensitive". Then no increase of the dependent variable will be observed for any of the subjects.

4. The subject in the experimental group is "sensitive", the corresponding subject in the yoked control group is "insensitive". Then an increase of the dependent variable is observed for the first subject, while no increase is observed for the second subject.

In three of the four cases the same effect is observed for both groups and only in the fourth case a higher increase is found in the experimental group. Altogether we get a higher increase of the cardiovascular reactivity in the experimental group which is the higher, the higher the percentage of pairs of type 4. According to this argumentation, an increased cardiovascular reactivity in the experimental group does by no means have to be an effect of "active coping", which means that subjects in the experimental group could avoid aversive tones by a good performance.

By the way, in the study [1] cited, actually an increased cardiovascular reactivity was reported for the experimental group, just as we predicted it under our assumptions.

c. Varying the time granted to solve a task varied the difficulty of the tasks. However, it is scarcely possible to perform a reasonable comparison of two respective groups which differ only with respect to the difficulty of the task: if, as it was the case here, the total time is fixed, the groups differ considerably with respect to the number of tasks. Here, the effects on the dependent variable could be explained, e.g., by greater learning gains for the subjects with shorter task duration because they obtained more feedback concerning their performance. An opposite effect could be due to the higher effort with respect to concentration, which is caused by the higher number of tasks.

If instead of the total time the total number of tasks is kept constant, a far higher total time must be assumed for the tasks with a higher permitted duration, by which, e.g., the effects of fatigue may be different for two corresponding groups. An opposite effect could be due to the fact that subjects with tasks which are granted a longer duration, get more positive feedback altogether, which may cause an increase in motivation.

For varying the difficulty of the tasks it would have been better to fix the duration of the tasks and to present all available tasks in a preliminary study to another sample of subjects from the same population. Then it would have been possible to form a subset of tasks, for which only for few subjects errors occurred and another subset of tasks of the same size where for many subjects errors occurred. Thus, the difficulty of the tasks could have been varied, while the total number and the total time of the tasks had been fixed.

4.13 King Psammetichos II wondered, to which people the first men on earth belonged. This was reduced to the question: Which is the oldest language of mankind? Consequently this question was tried to be answered by an isolation of new-born babies from any human language, record their first spoken word, and to identify the language to which this word belonged.

a. Even if it was possible to prove by means of the proposed procedure that children who do not learn to speak by other human beings, develop a language of their own, this would by no means prove that this language is the oldest language of mankind. This holds even, if the words produced by these children belong to a known language. This fact might be at most an indication that this language is particularly well adapted to the articulatory apparatus of the generations of people who lived at the respective time. However, this does not give any evidence with respect to a "natural" language of generations of people living a long time before. Therefore, the question concerning the oldest language of mankind cannot be answered by studying people of a given time period. It is only possible to ask the question whether there exists a "natural" language at the present time which is best adapted to the articulatory apparatus. This reduced problem is considered in the following.

b. Two possible outcomes of the study are conceivable if the kind of isolation of the children is varied in order to rule out alternative explanations with respect to the words produced by the children, e.g., that the children only try to imitate goats. First the childrens' first words might vary with the type of isolation. Then one can conclude that no uniform "natural" language of the present mankind exists, as assumed. Second, the same first words might always be produced irrespective of the kind of isolation. Then a causal conclusion is not possible because one cannot rule out that further kinds of isolation might exist which have not been considered and where the children would produce first words which belong to other known languages or which belong to no known language. Therefore, any study can have only the object to prove that no "natural" language exists.

c. For a study for proving that no "natural" language exists, several factors should be varied:

1. "Gender of the children" with the three levels "only boys", "only girls", and "half boys, half girls"
2. "Number of children" with the two levels "two children" and "four children"
3. "Ethnic descent of the children" with the three levels "Egyptian", "Phrygian", and "Hellenic"
4. "Animals" with the three levels "goats", "dogs", and "no sound from an animal is audible"
5. "Gender of the care-taker" with the two levels "male" and "female"
6. "Ethnic descent of the care-taker" with the three levels "Egyptian", "Phrygian", and "Hellenic"

Such a design would yield 3 × 2 × 3 × 3 × 2 × 3 = 324 different factor level combinations, i.e. there would be a good chance to detect actually existing differences. If, however, for all level combinations the same first words would be produced this would be strong evidence for a "natural" language even though a causal conclusion would not be admissible. To make obvious alternative explanations implausible, for each level combination the same number of care-takers should be used (e.g., only one care-taker). It is also advisable to use only mute care-takers.

4.14 In most cases the use of repeated measures as a control technique is said to have the advantage that a smaller sample size is needed, that the error variance is decreased, and that the expenditure is lower. However, these pretended advantages do not exist in reality because the outcomes of designs with repeated measures permit only then causal conclusions if additional control groups are being used, i.e. if the sample size and the expenditure are raised considerably. The reason for this are numerous alternative explanations such as history, maturation, testing, instrumentation, statistical regression, and selection. Further, a higher rate of experimental mortality has to be expected because of the longer duration of the studies and the higher stress of the subjects caused thereby. Even if the statistical dependence of the data, which is caused by repeated measures, is taken into account when the outcomes are being evaluated, additional assumptions are necessary whose validity is uncertain.

Answers to Chapter 5

5.1 Preliminary experiments serve as a check of to which extent the researcher has managed to realize the independent variables in the way intended and to measure the dependent variables with the desired reliability. These questions should be settled before a pilot study is performed. Dispensing with a pilot study would mean that the theoretically founded concept of a study with a minute schedule, a precise estimate of expenditure and costs, and a fixed frame of the study could be realized in practice without any modification. If this confidence in a theoretic concept is not justified, this might affect the untimely stop of a study, which has possibly caused considerable costs up to this point of time. To minimize the probability of a loss it is advisable to perform a well-planned pilot study before the start of the main study.

5.2 a. As a rule, one can expect that the experience gained when performing a pilot study, inevitably causes modifications in the proceeding of the main study. Though, sometimes these modifications may seem rather negligible, one cannot rule out that their implications on the outcome of the main study are nevertheless considerable.

b. As one tries to check all essential aspects of the main study in a "test-run" in pilot studies, despite a considerably reduced sample size, in most cases one dispenses with the randomization, to simulate the expected variability in a systematic way. If outcomes gained in this way would be enclosed into those of the main study this would mean a breaking of randomization whereby selection effects could no longer be ruled out.

Even if a randomization was performed in the pilot study it will not necessarily be compatible to the randomization performed for the main study.

c. Because the pilot study is performed before the main study, one cannot rule out that effects like, e.g., history can render the outcomes of both studies incompatible.

d. Often subjects are included into a pilot study whose "costs" are low because for some reason or other they can be easily recruited. The subjects in the main study, however, are mostly "naive" with respect to the study, i.e. they, among other things, neither know the experimenter nor have any pre-experiences. Therefore, one cannot assume that subjects which were recruited for the pilot study are comparable to the subjects in the main study in all respects.

e. The attitude of the experimenter towards the subjects in the pilot study, as a rule, differs from his or her attitude towards the subjects in the main study. The object of the pilot study is not to detect any effects but one only wants to make sure that the planned main study will be undisrupted. Therefore, the requirements with respect to the control of disturbance effects, e.g., of experimenter effects, are very low in a pilot study in comparison with the main study. Perhaps, this is one of the reasons why the outcomes of main studies very often differ considerably from the outcomes of preceding pilot studies.

All these considerations show that the outcomes of pilot and main studies cannot be considered as comparable. Therefore, in the interest of structural equality, one should dispense with pooling the data of both kinds of studies. Otherwise, biased results cannot be ruled out, even in spite of the small size of the pilot study.

5.3 If a pilot study has led to an essential modification of the study design, the original study cannot be realized as planned. In most cases one will find that each further pilot study causes further modifications of the study protocol because the "perfect" study design most probably does not exist. Therefore, at a certain stage of this proceeding one is compelled to hope that a further pilot study will not induce further important modifications of the study design and will start the main study. After the main study one will, as a rule, find that the original study design might have been improved with respect to one or more aspects. However, this would still have occurred after several further pilot studies. It is probably not possible to observe all complications, which may occur in the main study using a pilot study with a far smaller sample size.

It is thus obvious that the iterative process of finding an optimal study design via a sequence of pilot studies has to be stopped at some point of time to start the main study. Nevertheless one should keep the risk in mind that one takes if one starts the main study after considerable modifications of the study protocol without performing a further pilot study. In the worst case, one will find out after the start of the main study that it cannot be performed as planned and that due to obvious shortcomings the outcomes do not permit conclusive interpretations.

Answers to Chapter 6

6.1 Smoking behavior as well as the incidence of lung cancer were recorded for a large number of subjects to test the hypothesis that smoking increases the probability of the incidence of lung cancer. One might have observed far more patients with lung cancer among heavy smokers than among non-smokers, while for moderate smokers no distinct effect has been detected. As far as the direction of the relation between smoking and lung cancer is concerned, it is hardly plausible that lung cancer makes patients, who are suffering from it, increase smoking. This causal direction can be regarded implausible if, in addition, the smoking behavior of all patients is recorded before the assumed beginning of the disease, and one finds that an increase of smoking was not observed after the outbreak of the disease.

Due to the missing randomization, i.e. as it is not possible to randomly assign the property "heavy smoker" to subjects, one cannot conclude that an increased incidence is caused by heavy smoking. This is because one can neither rule out that the subsample of subjects with lung cancer does not only differ systematically with respect to smoking behavior but also with respect to other characteristics from the subsample of subjects without lung cancer, nor can one be sure that the true cause of the disease is not to be found among these other characteristics. One possible characteristic might be a gene which leads to addictive behavior as well as to lung cancer.

Therefore, one tries to find as many close relatives of each subject of the sample ("target subjects") as possible, preferably twins, who did not live together with the target subjects, at least since the latter started smoking. If a considerably smaller percentage of ill subjects is found among the non-smokers in the sample of relatives in comparison with the sample of smokers in the target sample this might be an evidence for the conclusion that smoking causes lung cancer, since the alternative explanations based on genetic factors and certain environmental factors would lose plausibility. A just as high or even higher percentage of lung cancer patients in the non-smoking subsample of relatives would rather be an indication for a genetic cause of lung cancer and would at least render smoking implausible as the sole cause of the disease.

Note that though it is possible to render certain alternative explanations implausible by considering additional samples (here: a sample of relatives) and additional measurements (here: degree of relationship, smoking history, case history) this does not permit true causal conclusions. The reason for this is that, e.g., genetic selection is only one of infinitely many possible sources of selection. Other possible factors of selection might be, e.g., the presence of noxious agents at the place of work, at home or in the neighborhood. Even if it was possible to form homogeneous samples with respect to these possible sources of selection other alternative explanations could not be ruled out.

6.2 Within-subjects designs should never be used to "save" subjects or patients or animals. Causal conclusions based on outcomes of within-subjects designs are only possible if the effects of more than one treatment and/or of more than one measurement can be cleanly separated from other effects by the use of additional control groups.

Within-subjects designs have to be used if, in particular, the effects of more than one measurement and/or of more than one treatment at the same subject are to be studied. It is important that a separate independent group of subjects is being used for each possible combination of measurements or treatments, respectively.

6.3 According to the definition, the precision of a design is the higher, the smaller the part of error variance, i.e. that part of variance observed in the outcomes which cannot be explained by the different treatment conditions. By this the precision of a design depends on the selection of the statistical procedure to be used, i.e. it depends also on the assumptions one is ready to accept as valid though these cannot really be tested.

One had better select a design with respect to that point of view that it permits to draw causal conclusions from the results, whose validity is not restricted to the case where certain model assumptions hold.

6.4 Since one cannot rule out that an initial treatment (e.g., A_1) just as an initial measurement has an effect on a second treatment (e.g., B_2) or on a second measurement, only the result of the comparison of the measurements immediately after the initial treatments (A_1 and B_1) in Figure 6.1 permits a causal interpretation. In case of the measurements after the treatment sequences A_1B_2 and B_1A_2 one cannot distinguish between the effects of the second treatment (e.g., B_2) and those of the initial treatment (e.g., A_1). It is not clear to what extent the initial treatment (e.g., A_1) has modified the outcome of the second treatment (e.g., B_2). By comparison of the outcomes after the second treatment (e.g., B_2) of one group and of the outcomes after the initial treatment (e.g., B_1) of the other group the total effect may be estimated. I.e. the rest effect of the initial treatment plus the modified effect of the second treatment is revealed. But even if no influence of the initial treatment (e.g., A_1) can be detected, the measurements after one treatment for both phases (e.g., measurements after B_1 and B_2) must not be pooled into one common sample. The reason for this is that the failure to detect an effect does not mean that this effect does not exist. Possibly, the sample sizes were not chosen large enough to make the detection of the effect possible.

6.5 If, in Figure 6.1, the scores after the treatments A_1 and A_2 or after B_1 and B_2, respectively, are being pooled, different kinds of wrong interpretations are possible, some of which will be discussed in the following.

Situation 1: Treatment A may have a positive effect of size *a* in relation to a control condition which is not considered here. Treatment B may also have a positive effect, which is, however, half as big as the effect of A (.5*a*). Further, treatment B might have the effect that after B has been applied no further treatment will be effective. In particular, the effect of B might be unchanged even after a subsequent application of A. If after A another treatment follows, the effect of this treatment is added to the effect of A.

When comparing A_1 with B_1 we find a superiority of treatment A with respect to treatment B because of $a > .5a$. If one assumes the sample sizes to be equal in both groups the average effect equals $.5(a + .5a) = .75a$ when the scores are pooled after A_1 and A_2, but $.5(.5a + 1.5a) = a$ when the scores are pooled after B_1 and B_2. This would yield the wrong impression of treatment B being superior.

In case of unequal sample sizes the direction of the results is the same.

Situation 2: Treatment A and treatment B may have the same positive effect of size a with respect to a control condition which is not considered here. If treatment B follows after treatment A, the two effects are added to ($2a$). However, if treatment A succeeds treatment B, A has no additional effect such that for the measurement after treatment A only an effect a due to treatment B is observed.

Comparing the outcomes of A_1 and B_1 no different effects of the treatments are found. If one assumes that the sample sizes are equal in both groups, the average effect if the scores are pooled after A_1 and A_2 is given by $.5(a + a) = a$, and if the scores are pooled after B_1 and B_2 by $.5(a + 2a) = 1.5a$. Again, the wrong impression of the superiority of treatment B arises.

In case of unequal sample sizes the direction of the results is the same.

Situation 3: In comparison with a control group, which is not considered here, treatment A may have a positive effect of size a and treatment B a negative effect ($-a$) of the same absolute size. If another treatment follows after treatment B, e.g., treatment A, the negative effect of treatment B is increased in such a way that after both treatments have been applied together, a negative effect of twice the size results, i.e. of ($-2a$). If another treatment follows after treatment A, e.g., treatment B, only treatment A is effective, i.e. after applying both treatments the effect a results.

By comparing the outcomes of A_1 and B_1 treatment A proves to be superior to treatment B because of $a > (-a)$. For the same sample sizes in both groups, the average effect if the scores are pooled after A_1 and A_2 ($.5(a - 2a) = -.5a$) is negative, while after a pooling of the scores after B_1 and B_2 ($.5(-a + a) = 0$) no difference with respect to the control condition is found. Again, the wrong impression of the superiority of treatment B results.

In case of unequal sample sizes, again, a superiority of treatment B over treatment A is found.

If here, the sample size for group A_1B_2 was more than twice as high as the sample size of group B_1A_2, positive average effects would result for both treatments after pooling. If the sample size for group A_1B_2 is larger than the sample size for group B_1A_2 but smaller than twice the sample size of group B_1A_2, after pooling a positive average effect results for B but a negative average effect for A. If the sample size for group A_1B_2 is smaller than that for group B_1A_2, negative average effects result for both treatments after pooling.

6.6 In multifactorial designs interactions are present if different effects of one independent variable with respect to direction and/or size are obtained, if the levels or level combinations of the remaining independent variables are varied at the same time. By averaging over all effects of the considered independent variable which result for all possible levels or level combinations, respectively, of the remaining independent variables, main effects are being revealed. These main effects may differ with respect to direction and/or size from all effects over which the average has been computed and, therefore, may yield an at least partially misleading interpretation of the effect of the independent variable in a given situation. E.g., a positive average effect of a drug can result though we know from the study outcomes that the drug has a negative effect in certain situations.

If a sample of patients is randomly split up into two subsamples, one of which gets a drug and the other one a placebo, the observed positive effect of the drug would be an average effect with respect to the population of patients for which the sample of patients is representative. The causal conclusion that the drug has an effect, which is superior to the effect of the placebo, with respect to the respective population is admissible. Nevertheless, this does not mean that there might not be a subpopulation for which the drug does not have any or even a harmful effect. In contrast to the multifactorial design this design would not permit the identification of conditions for such different effects. To do so, it is necessary to know corresponding subpopulations, to select a representative sample from each of these subpopulations, to split this up randomly into two subsamples, and to apply the drug or the placebo, respectively, to these subsamples. Because there exists, in general, a very large number of possible subpopulations, this kind of proceeding cannot be used in practice.

In contrast to the proceeding described above in case of interactions in a multifactorial design, a randomization yields average effects which arise in a random and nonsystematic way. In this case the probability that a non-existing effect is detected can be kept below a given small threshold, which is usually called α, by using statistical procedures. If, however, effects in a multifactorial design are averaged in the presence of interactions in many cases wrong conclusions are drawn with respect to the direction and/or size of the effects, though this could be avoided.

6.7 Obviously, the subjects were not randomly assigned to the four considered samples. Therefore, one cannot rule out that the four samples do not only differ with respect to the two recorded characteristics (presence or absence of "obesity" or "binge eating") but also with respect to other characteristics which have not been recorded. Thus, any observed group differences must by no means be associated with the two considered characteristics but may be due to totally different causes. Even before the study it was obvious that outcomes permitting causal interpretations would not result.

6.8 a. Two animals received the sequence $A_1P_2P_3A_4$, the two remaining animals the sequence $P_1A_2A_3P_4$, where "A" denotes the atropine condition and "P" the placebo condition. A measurement after A_1 is only influenced by A_1, but a measurement after A_4 is influenced by A_1, P_2, P_3, and A_4 in this order of conditions. By analogy, a measurement after A_2 is not only influenced by A_2 but also by P_1, while a measurement after A_3 is influenced by P_1, A_2, and A_3

in this order. Therefore, the measurements after A_1, A_2, A_3, and A_4 cannot be compared with each other due to the preceding treatment conditions and it is by no means permitted to consider them as equivalent. If, however, they are treated as equivalent, no causal conclusions are possible due to the non-separable effects of the different treatments, because any imaginable outcome can be explained in different ways.

b. Even if only one experimental condition had been applied to each of 16 animals, i.e. if 8 atropine and 8 placebo animals had been considered, the comparison of the measurements before and after eating would not have permitted a causal interpretation. The detected "effect" for the atropine animals might be solely due to intermediate events. The non-significant result for the placebo animals can by no means be interpreted in that way that here no effect occurred. The only comparison, which would have permitted an interpretation, would have been the comparison of the scores for the two samples after eating.

c. In case of four treatments, pretests, and posttests, and two dependent variables, 16 measurements are obtained for each animal which are probably statistically dependent to a high degree. If more than one of these scores is used in the same statistical test (and in this study four scores of the same animal were used in each test!) a reasonable interpretation of statistical results is no longer possible, except if very implausible additional assumptions are made, the validity of which cannot be really ascertained.

All in all, we can state: Even before the study started, it was obvious that due to the poor planning the study could not yield outcomes which permit a sound interpretation. Only a comparison of the scores at the end of the first session would be admissible with two animals in each of the two groups. However, these sample sizes are far too small for a reasonable comparison, in particular, as one of the four animals, the domestic pig, most probably cannot be compared to the three other animals.

6.9 a. The treatment sequence F_1P_2 was applied to four subjects, and the treatment sequence P_1F_2 to the four remaining subjects. Here, "F" denotes the condition with meal and "P" the condition without meal. A measurement after F_1 is only influenced by F_1, but a measurement after F_2 by P_1 and F_2 in this order of conditions. By analogy, a measurement after P_1 is influenced only by P_1, but a measurement after P_2 by F_1 and P_2 in this order of conditions. From this follows that the measurements after F_1 and F_2 (and likewise the measurements after P_1 and P_2) cannot be compared with each other and cannot be considered as equivalent. However, if this equivalence is assumed, as in the cited study, several alternative explanations exist for each outcome.

b. Even if 16 subjects, each with only one kind of treatment, had been considered, i.e. 8 subjects with meal and 8 subjects without meal, the comparisons of the 8 pretest and 8 posttest values for one condition would not have permitted a causal interpretation. Any effects might have been solely due to intermediate events. This would not have been ruled out either, if under condition F a significant result, but under condition P a non-significant result had been obtained, because a non-significant result does not mean that no effect is present. Only a comparison of the samples based on the scores recorded after applying the respective conditions would have permitted a conclusive interpretation. A comparison of the pretest values would have been superfluous for a causal interpretation if an appropriate randomization had been performed.

c. In case of two treatments, respective pretests and posttests, and two dependent variables 8 measurements are obtained for each subject, which are probably statistically dependent to a high degree. Because two of these dependent values were used for each subject in each statistical test, a reasonable interpretation of the statistical results is not possible without additional assumptions, which are neither plausible nor can really be checked.

As a result we can state: Only a comparison of the respective four measurements after F_1 or P_1 might have made sense. The measurements before and after F_2 and P_2 do not permit conclusive interpretations. Because these latter measurements cause stress for the subjects it would have been better to dispense with them. Because it was obvious that a respective sample size of 4 subjects was too small to detect effects in view of the expected variances, an appropriate experimental design would have required that in each sample at least 8 subjects had been used, i.e. altogether 16 subjects, which always had been measured for only one condition. This would have meant half of the stress for each subject, compared to the cited study, while the total expenditure and the total costs would have been the same.

6.10 Because the levels of the independent variable (migraine vs. no migraine) could not be randomly assigned to the two samples, one cannot rule out that the groups differ also with respect to other characteristics. Therefore, it is possible that any differences with respect to specific psychophysiological responses are not related to the respective characteristic. Additionally, one cannot rule out that such selection effects prevent that effects that actually exist are detected.

In order to render at least the most obvious alternative explanations implausible, it would have been advisable to perform a careful matching, e.g. with respect to age, gender, and educational standard. Because such a matching has not been performed it is not clear whether the high differences in age led to the observed effects or to a non-detection of effects, which actually exist. In view of the way in which subjects were recruited for this study it was obvious, even before the study, that one would not obtain results permitting a conclusive interpretation.

6.11 Because the levels of the independent variable (migraine vs. no migraine) could not be randomly assigned to the two samples, selection effects cannot be ruled out. The description of the two samples already shows that they differed considerably not only with respect to the disease but also with respect to the educational standard. Any effects, which might be found, can be explained by this difference alone. On the other hand, it might also be possible that effects that actually exist cannot be detected due to the different compositions of the samples with respect to educational standard. To render such obvious alternative explanations implausible, a matching with respect to, e.g., age, gender, and educational standard would have been advisable. In view of the way in which subjects were recruited for the study it was clear from the beginning that no interpretable outcomes could be obtained.

Answers to Chapter 7

7.1 If more than one measurement is recorded for each subject, effects of the measuring process on subsequent measurements cannot be ruled out. Likewise one cannot rule out that a treatment has an effect on measurements which are only recorded after succeeding treatments if we are in a situation in which a sequence of more than one treatment is applied to each subject. In order to estimate the pure effect of a treatment, i.e. an effect which is not modified due to pre-treatments or pretests, additional subjects are needed, to which the respective treatment is being applied before any other treatment or measurement. As such an independent sample of subjects is required for each treatment, the sample size for a repeated-measures design which permits a causal interpretation has to be at least as large as that for a corresponding design without repeated measures.

7.2 One-factor designs have the following advantages as opposed to multifactorial designs:

1. They permit a more simple interpretation because the possible existence of differential effects can be ignored.
2. They require, as a rule, fewer groups of subjects, i.e. the total sample size is also smaller.

A disadvantage of one-factor designs is that the joint effect of several factors which affect a dependent variable simultaneously cannot be investigated.

7.3 If only one treatment group is being used, it is not possible to prove the existence of the effect of the treatment since one cannot rule out that the effect might also have taken place without the treatment. This deficiency is also present if a difference between the pretest and the posttest measurement of the dependent variable is observed. Such a difference might have been caused, e.g., by an effect of the pretest or by intermediate events, which are in no relation to the treatment.

If in addition to the treatment group, a control group is being used, for which the only difference in comparison to the treatment group is that the subjects do not receive a treatment or the treatment is replaced by a control condition, any difference between the two groups has to be due to a difference in the two conditions.

With two or more treatment groups it is certainly possible to detect differences between the treatments, but nothing can be said about the "absolute" effect of the treatments, if no control group is being used. However, if control groups are considered as a specific kind of treatment groups, one could argue that even if control groups are being used, one can only detect "relative" effects, but no "absolute" effects.

7.4 The effect of "noise with meaning" on "concentration" is to be investigated. Here, "noise with meaning" is operationalized by a radio programme on an art exhibition given in the subjects' mother tongue and is applied via headphones like in the control conditions. The dependent variable "concentration" is operationalized in that way that subjects have to press on a key each time they perceive one of two symbols in a random sequence of symbols which move from the right to the left of a window on a screen. The actual dependent variable might be, e.g., the number of correctly identified symbols or the number of wrongly identified symbols.

A first control group does not get any noise; a second control group gets white noise; a third control group gets traffic noise; a fourth control group gets a piece of modern classical music, e.g., "Spiral" by Karlheinz Stockhausen. Each of the four control groups is related to another aspect of the opposite of "noise with meaning". One hopes that it is possible to isolate that or those aspects of "noise with meaning" which are responsible for a decrease in the performance of the concentration task by comparing these control groups to the experimental group. Additional control groups might get texts in an unknown language or songs in an unknown language, where on the one hand songs with well-known melodies and on the other hand songs with unknown melodies might be chosen.

7.5 The relative effect of a treatment A with respect to a treatment B can be measured by randomly assigning subjects to the two treatments and by considering the difference in the effects of the two treatments on a dependent variable. If both treatments have the same effect, no relative effect will be obtained.

A treatment A might show a relative effect with respect to a treatment B though it has no effect on the dependent variable. In this case the treatment A has no absolute effect. In order to detect an absolute effect of a treatment, the treatment group is compared to a control group. The conclusion that an effect found in this way is de facto an absolute effect, can only be drawn if one assumes that the control condition has had no effect on the dependent variable.

7.6 The $m = 2$ treatments T1 and T2 may correspond to two different drugs which lower the blood pressure. The $k = 2$ control conditions C1 and C2 may, on the one hand, correspond to the application of a placebo, on the other hand to the lapse of time without any intervention.

7.7 The relation between the increase of the dose of the drug and the lowering of the body weight of mice is to be investigated for an anorectic. The functional relationship between the dose as independent variable and the change of weight as dependent variable is called dose-response curve. In many cases, not the functional relationship itself but its graphic representation is denoted a dose-response curve. To obtain a dose-response curve, at first the doses are fixed according to the points of view, as discussed in Section 2.2. We thus get, e.g., 8 different doses. A sample of 80 mice is randomly split into 8 subsamples, each with 10 mice, which correspond to the 8 doses of the drug. The weight of each mouse is being measured at the same time of the day, before the mouse gets the dose to which it was assigned. After two days the weight is measured a second time at the same hour. Before and during the study the mice have free access to water and food and are all kept under exactly the same conditions, with the exception of the dose of the drug which differs between groups. The difference between posttest and pretest score for each mouse is taken as a measure of weight change. By means of a statistical nonlinear regression procedure the functional relationship between dose and weight change can be estimated. A first impression of this function is obtained by inserting points into a coordinate-system where the x-values correspond to the doses and the y-values to the corresponding means of weight changes for 10 mice.

7.8 A new kind of packing is sought for a chocolate product. A first factor has the two levels "circular" and "heart-shaped". A second factor has the two levels "red" and "blue". A sample of 40 subjects is randomly split into four subsamples each with 10 subjects. To each subsample one of the four possible combinations (red circle, blue circle, red heart, blue heart) is assigned at random. The subjects have to rate the attractiveness of the combinations on a 10 point rating scale. One might find that the average attractiveness of a red packing is higher than that of a blue one, but that the difference between the two colors is larger for heart-shaped packing than for circular one. This outcome indicates an interaction between color and shape of the packing. Such an apparent interaction would not be observed if the difference between the colors red and blue were the same for circular and heart-shaped packings.

7.9 The design in Figure 7.15 consists of two one-factor subdesigns for which altogether six samples are needed. Relative treatment effects can be detected by comparing A1 with A2 or B1 with B2, and absolute treatment effects by comparing A1 with A0, A2 with A0, B1 with B0, and B2 with B0. This design has the advantage that it permits to draw several easy to interpret and easy to formulate causal conclusions on the basis of only few samples. A possible disadvantage is that nothing can be said about the simultaneous effect of both factors.

The design in Figure 7.13 requires five samples and has the advantage to permit conclusions about the joint effect of both factors. A disadvantage is that only relative effects of two treatments can be studied. With the fifth sample (control condition C) one can decide for each of the four factor level combinations whether it has had any effect or not.

The design in Figure 7.12 requires nine samples which at the same time is its most important disadvantage. A further disadvantage is that due to the many possible comparisons of samples a corresponding large number of differential effects may be detected which makes it difficult to formulate general conclusions about the joint effect of both factors. An advantage of this design is that as well relative as absolute effects of both factors can be studied and in addition those effects which are caused by both factors together.

7.10 Treatment A may be a first drug with the doses 0 mg (saline) corresponding to A0, 1 mg corresponding to A1, and 2 mg corresponding to A2. Treatment B is a second drug with the doses 1 mg corresponding to B1, 2 mg corresponding to B2, 4 mg corresponding to B3, and 8 mg corresponding to B4. Further, under the control condition B01 nothing is given, under the control condition B02 water, and under the control condition B03 saline. This design is depicted in Figure 7.22.

	B01	B02	B03	B1	B2	B3	B4
A0	A0B01	A0B02	A0B03	A0B1	A0B2	A0B3	A0B4
A1	A1B01	A1B02	A1B03	A1B1	A1B2	A1B3	A1B4
A2	A2B01	A2B02	A2B03	A2B1	A2B2	A2B3	A2B4

Figure 7.22: Two-factor design with one control and two treatment conditions for factor A and three control and four treatment conditions for factor B

7.11 Because there are 2 × 3 × 4 = 24 level combinations (cf. Figure 7.23), 24 samples are needed. From 24 level combinations 24 × (24 – 1) / 2 = 276 pairs can be considered and correspondingly many differential comparisons by analogy to Figure 7.19. If pairs are formed from these comparisons for comparing the differences with another, we get 276 × (276 – 1) / 2 = 37950 pairs of pairs.

1. A1B1C1	5. A1B2C1	9. A1B3C1	13. A2B1C1	17. A2B2C1	21. A2B3C1
2. A1B1C2	6. A1B2C2	10. A1B3C2	14. A2B1C2	18. A2B2C2	22. A2B3C2
3. A1B1C3	7. A1B2C3	11. A1B3C3	15. A2B1C3	19. A2B2C3	23. A2B3C3
4. A1B1C4	8. A1B2C4	12. A1B3C4	16. A2B1C4	20. A2B2C4	24. A2B3C4

Figure 7.23: Level combinations of a three-factor design with two levels (A1, A2) of factor A, three levels (B1, B2, B3) of factor B, and 4 levels (C1, C2, C3, C4) of factor C

7.12 For a four-factor design, where each factor has two levels, 2 × 2 × 2 × 2 = 16 level combinations result (cf. Figure 7.24). Thus, 16 samples are needed. From 16 level combinations 16 × (16 – 1) / 2 = 120 pairs can be formed and the same number of differential comparisons is possible.

1. A1B1C1D1	5. A1B2C1D1	9. A2B1C1D1	13. A2B2C1D1
2. A1B1C1D2	6. A1B2C1D2	10. A2B1C1D2	14. A2B2C1D2
3. A1B1C2D1	7. A1B2C2D1	11. A2B1C2D1	15. A2B2C2D1
4. A1B1C2D2	8. A1B2C2D2	12. A2B1C2D2	16. A2B2C2D2

Figure 7.24: Level combinations of a four-factor design with two levels (A1, A2) of factor A, two levels (B1, B2) of factor B, two levels (C1, C2) of factor C, and two levels (D1, D2) of factor D

7.13 For the design in Question 7.11 the number of possible pairwise comparisons is given by 24 × (24 – 1) / 2 = 276 and for the design in Question 7.12 by 16 × (16 – 1) / 2 = 120.

7.14 A tendency for a main effect of factor A can be found in Figure 7.23 if the average of all scores for the level combinations No. 1 to No. 12 differs distinctly from the average of all scores for the level combinations No. 13 to No. 24.

For the detection of a main effect of factor B the three averages over the level combinations No. 1 – No. 4, No. 13 – No. 16 and No. 5 – No. 8, No. 17 – No. 20 and No. 9 – No. 12, No. 21 – No. 24, respectively, are compared.

For the detection of a main effect of factor C the four averages over the level combinations No. 1, No. 5, No. 9, No. 13, No. 17, No. 21 and No. 2, No. 6, No. 10, No. 14, No. 18, No. 22 and No. 3, No. 7, No. 11, No. 15, No. 19, No. 23 and No. 4, No. 8, No. 12, No. 16, No. 20, No. 24, respectively, are compared.

In order to detect a first-order interaction between factors A and B, first of all six averages are computed with respect to the scores for the factor level combinations No. 1 to No. 4 and No. 5 to No. 8 and No. 9 to No. 12 and No. 13 to No. 16 and No. 17 to No. 20 and No. 21 to No. 24, respectively. Then the three differences average 4 – average 1, average 5 – average 2, average 6 – average 3 are computed. If these differences differ considerably, there might be an interaction of the form A × B.

In order to detect a first-order interaction between factor A and factor C, first eight averages are computed with respect to the scores for the factor level combinations No. 1, No. 5, No. 9 and No. 2, No. 6, No. 10 and No. 3, No. 7, No. 11 and No. 4, No. 8, No. 12 and No. 13, No. 17, No. 21 and No. 14, No. 18, No. 22 and No. 15, No. 19, No. 23 and No. 16, No. 20, No. 24, respectively. Then the four differences average 5 – average 1, average 6 – average 2, average 7 – average 3, average 8 – average 4 are computed. If these differences differ considerably there is an indication for the existence of an interaction of the form A × C.

To detect a first-order interaction between factor B and factor C, first twelve averages are computed with respect to the scores for the factor level combinations No. 1, No. 13 and No. 2, No. 14 and No. 3, No. 15 and No. 4, No. 16 and No. 5, No. 17 and No. 6, No. 18 and No. 7, No. 19 and No. 8, No. 20 and No. 9, No. 21 and No. 10, No. 22 and No. 11, No. 23 and No. 12, No. 24, respectively. Then the four differences average 5 – average 1, average 6 – average 2, average 7 – average 3, average 8 – average 4 are computed. If these differences differ considerably an interaction of the form B × C might exist. Likewise the four differences average 9 – average 5, average 10 – average 6, average 11 – average 7, average 12 – average 8 can be considered. Again, if these differences differ considerably this is an indication of the possible existence of an interaction of the form B × C.

In order to detect a second order interaction between the factors A, B, and C, first of all 24 averages are computed with respect to the scores for the factor level combinations No. 1 to No. 24 in Figure 7.23. With respect to these averages, we may consider, e.g., the difference of differences

(A2B1C1 – A1B1C1) – (A2B2C1 – A1B2C1)

and compare this to

(A2B1C2 –A1B1C2) – (A2B2C2 –A1B2C2).

If the results differ considerably this possibly reveals an interaction of the form A × B × C. Other comparisons with the same object would result if we considered in the comparison above instead of the levels (B1, B2) for factor B the levels (B1, B3) or (B2, B3), and instead of the levels (C1, C2) for factor C the levels (C1, C3), (C1, C4), (C2, C3), (C2, C4) or (C3, C4). Altogether, we would consider 3 × 6 = 18 such comparisons. Only one of them is explicitly given above.

7.15 For four factors A, B, C, and D, each with two levels, the following effects are possible:

a. Four main effects: A, B, C and D
b. Six first-order interactions: A × B, A × C, A × D, B × C, B × D, C × D
c. Four second-order interactions: A × B × C, A × B × D, A × C × D, B × C × D
d. One third-order interaction: A × B × C × D

7.16 For the design in Question 7.11 result with 3 possible main effects and 3 + 1 = 4 possible interactions altogether 7 possible tests in an analysis of variance.

For the design in Question 7.12 result with 4 possible main effects and 6 + 4 + 1 = 11 possible interactions altogether 15 possible tests in an analysis of variance.

7.17 For the design in Question 7.12 altogether 15 different tests are possible which can be tested in an analysis of variance. Since each of these effects can be present or absent, altogether $2^{15} = 32768$ different effect patterns are possible.

Answers to Chapter 8

8.1 a. The state of health of a sample of patients after an operation is checked in regular time intervals. The question whether the measurements are reactive arises, i.e. whether the measuring procedure itself has an influence on the following measurements. If the question is to be answered by means of an experiment, a repeated-measures design has to be applied.

b. In order to rate the state of health of patients after an operation, several dependent variables are recorded. Here, one cannot rule out that different measuring procedures influence each other. If this is investigated in an experiment and if one assumes that the measurements are not performed in parallel but in sequence, a repeated-measures design has to be used.

8.2 It is advisable in the following situations to record more than one measurement after a treatment:

a. The effect of the treatment does not set in at once but only after a time interval of unknown duration, which might be different for each subject.

b. The duration of the treatment effect is not known and possibly different for each subject.

c. The height of the treatment effect is not constant in time, and, e.g., the time and the height of the maximum effect should be recorded for each subject.

8.3 a. If one cannot rule out that measurements are reactive, succeeding measurements may be influenced in a different way for different numbers or different time patterns of pretests and posttests. Observed group differences might be thus caused by the effects of reactivity and not by different treatment effects.

b. If measurements are not reactive, different numbers or different time patterns of the pretests can yield different baseline scores. In an analogous situation, the posttests might measure completely different aspects of the treatments, e.g., early and late effects. In both cases, non-existent group differences might be "detected". In the same way it may happen that due to the absence of measurements in a certain time interval for a particular group, distinct group differences cannot be perceived.

8.4 a. One wants to prove for a particular drug that, in principle, it is an effective anorectic. In a first stage, the conditions under which the drug is particularly efficient or under which it does not have any effect at all, respectively, are not investigated. The study is to be performed with healthy subjects with normal weight and one fears that the effect of the drug may be so low under these conditions that it cannot be detected due to the high variability of the subjects with respect to weight. Further, it is not clear whether an absolute dose or a weight-dependent dose of the drug should be used. Therefore, one requires the subjects to be healthy, male between 25 and 27 years, with a weight between 74 kg and 77 kg and a height between 180 cm and 185 cm as inclusion criteria for the study. For such a relatively homogeneous sample of subjects which was selected on the basis of pretests it is not necessary to distinguish between absolute and weight-dependent doses of the drug. Most probably, only a small sample can be recruited under these restrictions which then still has to be randomly split up into a placebo and a drug group. Because of the homogeneity of the sample the error variance is probably so small that also small effects of the drug can be detected.

b. In case of therapies, pretests are necessary to give appropriate diagnoses. Therefore, it is advisable to study possible effects of such pretests on the posttests by means of an experiment.

8.5 a. The duration of periods without treatment or control condition and without measurement should be chosen so long that direct transfer effects are implausible, except in the case of irreversible effects.

b. However, the duration of such periods should not be chosen to be too long because the subjects under study underlie changes due to maturation, history, and similar intermediate events. These effects might interact with the treatment effects. An example might be an accelerated maturation caused by a drug.

c. The periods should be chosen such that the next intervention of the experimenter is performed at a point of time which is comparable to the point of time at which the preceding intervention took place. One had better not, e.g., to start one intervention at 8:00 a.m. and the next one at 8:00 p.m. Here, an interaction between the time of intervention and circadian rhythms could not be ruled out.

8.6 a. In designs without repeated measures the dropout rate can be used as a dependent variable to measure a possible effect of the independent variable. E.g., a high dropout rate for a certain therapy, in comparison with other therapies, may be an indicator of aversive effects of this therapy. For repeated-measures designs the dropout rate is less suited as dependent variable because the time of the dropout of a subject cannot necessarily be set equal with the point of time when the cause was effective. If, e.g., an animal dies during the performance of a repeated-measures design, it may be difficult to find out whether one of the preceding treatments and, if so, which of these treatments led to the animal's death.

b. Repeated-measures designs are by definition more time-consuming for the participating subjects than designs for the same problem but without repeated measures. Therefore, the probability of the occurrence of dropouts increases with the complexity of the design. The resulting loss of data will, in any case, cause interpretational problems. To replace the lost data by the data of newly recruited subjects is problematic because of selection effects due to the incomplete randomization. Selection effects can also result if it is tried to compensate experimental mortality by larger sample sizes. In this case, it is by no means guaranteed that the dropouts form a random sample of the initial sample, so that a systematic bias cannot be ruled out.

8.7 a. After-effects are not controlled as long as they do not concern solely the immediately succeeding condition. Here, not only an outlasting effect of an experimental condition is possible but also accumulating effects.

b. Different responses to the conditions T1, T2, and C might have been due to subjects being able to discriminate the conditions in a way which has not been taken into consideration by the experimenter.

c. If no double-blind study is used, experimenter effects cannot be ruled out. These may also occur in a double-blind study if the experimenter can discriminate the conditions on the basis of the observed effects during the study.

d. Since the chronological order of the experimental conditions is fixed in advance one cannot rule out that a confounding with cyclic trends occurs, e.g., with the circadian rhythms of the subjects. History (cf. Section 3.2.1), i.e. effects from outside the experimental situation, which are not controlled by the experimenter, can also yield false interpretations. If, e.g., the laboratory is located under the flightpath of a nearby airport, and if the application times for treatment T1 are chosen unintentionally such that just at these times aeroplanes fly over the laboratory though this is not the case for conditions T2 and C, a causal conclusion can be made difficult.

8.8 The extension of the Solomon four group design to two treatments T1 and T2 is depicted in Figure 8.26. A comparison of M_2 and $M_2^{(1)}$ or of $M_2^{(2)}$ with $M_2^{(3)}$, respectively, shows whether an effect of treatment T1 is present. If two different effects result, this shows that the pretest M_1 has also an influence on the posttest M_2. The effect of the pretest M_1 can be separated by comparing $M_2^{(3)}$ with $M_2^{(1)}$ or $M_2^{(2)}$ with M_2, respectively. If here two different effects are found, the presence or absence of treatment T1 modifies the effect of the pretest.

	Point of Time		
Group	t_1	t_2	t_3
1	M_1	T1	M_2
2	$M_1^{(1)}$		$M_2^{(1)}$
3		T1	$M_2^{(2)}$
4			$M_2^{(3)}$
5	$M_1^{(4)}$	T2	$M_2^{(4)}$
6		T2	$M_2^{(5)}$

Figure 8.26: Extension of the Solomon design from Figure 8.19 to two treatments T1 and T2 and six groups

A comparison of $M_2^{(4)}$ with $M_2^{(1)}$ or of $M_2^{(5)}$ with $M_2^{(3)}$ shows whether an effect of treatment T2 exists. In case of two different effects, it can be concluded that the pretest $M_1^{(4)}$ has an influence on the posttest $M_2^{(4)}$. The effect of the pretest $M_1^{(4)}$ results from the comparisons $M_2^{(3)}$ with $M_2^{(1)}$ or $M_2^{(5)}$ with $M_2^{(4)}$. If here two different effects are found, then the presence or absence of treatment T2 modifies the effect of the pretest.

Finally, one can conclude whether the treatments T1 and T2 have different effects by comparing $M_2^{(2)}$ with $M_2^{(5)}$ or M_2 with $M_2^{(4)}$, respectively. If both comparisons yield different effect differences for the treatments T1 and T2, it can be concluded that the pretest affects the posttest in a different way if different succeeding treatments are used.

8.9 The idea that each subject serves as its own control if repeated-measures designs are used is so attractive because this technique takes advantage of the fact that the error variance is reduced by that part of the variance which is caused by intersubject differences. If the variance is reduced in this way it is much easier to detect existing effects, i.e. to draw causal conclusions.

Unfortunately, this idea is based on a model of measurement error which is too simple to be accepted in practice. It is acted as if a sample of independent subjects is replaced by another sample of independent and nearby identical subjects, e.g., animals from one litter or even better subjects generated by cloning. While the identity of the subjects can be assumed in case of repeated measures, the assumption of the independence of measurements, among others, is not plausible because of possible transfer effects of measurements and experimental conditions on successive measurements. In particular, if a positive correlation of the measurements recorded for one subject results due to this kind of effects, an underestimation of the real variance of the measurements has to be expected, with the consequence that non-existing effects are "detected". In practice, it is rather improbable that a negative correlation results, such that the variance would be overestimated and existing effects might not be detected.

Thus, one problem when repeated-measures designs are being used as opposed to designs with independent samples is that, though the influence of inter-subject variance is reduced, the assumption of the independence of measurements is no longer valid, and this assumption is essential for statistical conclusion validity.

8.10 Assume that a patient is rated on a rating scale with 10 ordered categories where the rating 1 means that the patient is so ill that he or she needs immediate medical care to survive, while the rating 10 means that the patient is healthy. If a doctor has to rate the state of health of always new patients it is possible that he or she experiences a change with respect to the interpretation of single categories, i.e. the calibration of the measuring instrument is changed. If, e.g., the last 100 patients had a rather good state of health, the doctor may be inclined to discriminate patients in this range better than patients with a bad state of health. This might have the consequence that category 5 is used for patients to which at the beginning of the study the category 6 or 7 was assigned. The opposite effect that category 5 is used for patients which were rated into category 3 or 4 at the beginning of the study could occur if the last 100 patients belonged to the more ill ones.

Another effect might be that the doctor in the course of the study becomes more and more sure in his or her judgement of patients by means of the scale. If at the beginning of the study two patients with a similar state of health may sometimes have been assigned to two different categories this becomes rare in the course of time, i.e. the reliability of the measuring instrument is increased.

8.11 The nine compounds of Figure 8.23 may be coded in the following way by Latin capital letters:
A = 4/60, B = 0/120, C = 2/60, D = 4/120, E = 0/60, F = 4/0, G = 2/120, H = 0/0, I = 2/0.

Using incomplete counterbalancing, a design for the nine compounds is depicted in Figure 8.27, where the 18 rows of the design correspond to the 18 subjects.

ABCDEFGHI
BCEGDIFAH
CDFAHGIEB
DHABFECIG
EGBICHDFA
FIHEBDAGC
GFICABHDE
HEGFIABCD
IADHGCEBF
DIAGHBECF
IAHEGFBDC
AGBDCEFHI
DCGIBHAFE
HEIFACGBD
BFCHIGDEA
EBFADICGH
CHEBFDIAG
FDGCEAHIB

Figure 8.27: Design for nine treatments and 18 subjects where each treatment occurs in each timely position exactly twice (incomplete counterbalancing)

8.12 In case of k treatments T1, ..., Tk and one control condition C the number of groups needed for a design as in Figure 8.16 is given by the formula

$$1 + k - k! + k\sum_{i=1}^{k} \frac{k!}{i!}$$

with

$k! = 1 \times 2 \times ... \times k.$

We give here no proof for this formula. For $k = 1$ we obtain

$$1+1-1!+1\times\frac{1!}{1!}=1+1-1+1\times1=2,$$

because 2 groups are to be considered with T1 and C. For $k = 2$ we obtain

$$1+2-2!+2\left(\frac{2!}{1!}+\frac{2!}{2!}\right)=1+2-2+2(2+1)=7,$$

which corresponds to the 7 groups in Figure 8.16.

In Figure 8.24a, b $k = 9$ compounds are considered. Thus, here

$$1+9-9!+9\left(\frac{9!}{1!}+\frac{9!}{2!}+\frac{9!}{3!}+\frac{9!}{4!}+\frac{9!}{5!}+\frac{9!}{6!}+\frac{9!}{7!}+\frac{9!}{8!}+\frac{9!}{9!}\right)$$

$$= 1 + 9 - 362880 + 9\ (362880 + 181440 + 60480 + 15120 + 3024 + 504 + 72 + 9 + 1)$$

$$= 5248900$$

groups have to be considered. Due to this large number of groups, this design cannot be used in practice.

8.13 In Figure 8.28 a, b, c, d, e the two-factor design with repeated measures only on the first factor "vitamin A", mentioned in Section 8.2, is depicted. Bear in mind that the 18 subjects of the figure should be a random arrangement of the 18 subjects of the sample.

Since both factors are effective simultaneously and because three pretests and three posttests are performed with each subject it is not quite correct to call the design described in Section 8.2 and depicted in Figure 8.28 a, b, c, d, e a design with repeated measures on the first factor. One had better call it a design with multiple treatments on the first factor.

Day	1	2	3	4	5	6	7
Subject1	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Subject2	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Subject3	2/0	2/0	2/0	2/0	2/0	2/0	2/0
Subject4	2/0	2/0	2/0	2/0	2/0	2/0	2/0
Subject5	4/0	4/0	4/0	4/0	4/0	4/0	4/0
Subject6	4/0	4/0	4/0	4/0	4/0	4/0	4/0
Subject7	0/60	0/60	0/60	0/60	0/60	0/60	0/60
Subject8	0/60	0/60	0/60	0/60	0/60	0/60	0/60
Subject9	2/60	2/60	2/60	2/60	2/60	2/60	2/60
Subject10	2/60	2/60	2/60	2/60	2/60	2/60	2/60
Subject11	4/60	4/60	4/60	4/60	4/60	4/60	4/60
Subject12	4/60	4/60	4/60	4/60	4/60	4/60	4/60
Subject13	0/120	0/120	0/120	0/120	0/120	0/120	0/120
Subject14	0/120	0/120	0/120	0/120	0/120	0/120	0/120
Subject15	2/120	2/120	2/120	2/120	2/120	2/120	2/120

Subject16	2/120	2/120	2/120	2/120	2/120	2/120	2/120
Subject17	4/120	4/120	4/120	4/120	4/120	4/120	4/120
Subject18	4/120	4/120	4/120	4/120	4/120	4/120	4/120
Score	B						

Figure 8.28a: Schedule of a two-factor repeated-measures design with repeated measures on the first factor, with the nine level combinations from Figure 8.23, for the first week, with pretests (B)

Day	8	9	10	11	12	13	14
Subject1							
Subject2							
Subject3							
Subject4							
Subject5							
Subject6							
Subject7							
Subject8							
Subject9							
Subject10							
Subject11							
Subject12							
Subject13							
Subject14							
Subject15							
Subject16							
Subject17							
Subject18							
Score	A						

Figure 8.28b: Schedule of a two-factor repeated-measures design with repeated measures on the first factor, with the nine level combinations from Figure 8.23, for the second week, with posttests (A)

Day	15	16	17	18	19	20	21
Subject1	2/0	2/0	2/0	2/0	2/0	2/0	2/0
Subject2	4/0	4/0	4/0	4/0	4/0	4/0	4/0
Subject3	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Subject4	4/0	4/0	4/0	4/0	4/0	4/0	4/0
Subject5	0/0	0/0	0/0	0/0	0/0	0/0	0/0
Subject6	2/0	2/0	2/0	2/0	2/0	2/0	2/0
Subject7	2/60	2/60	2/60	2/60	2/60	2/60	2/60
Subject8	4/60	4/60	4/60	4/60	4/60	4/60	4/60
Subject9	0/60	0/60	0/60	0/60	0/60	0/60	0/60
Subject10	4/60	4/60	4/60	4/60	4/60	4/60	4/60
Subject11	0/60	0/60	0/60	0/60	0/60	0/60	0/60
Subject12	2/60	2/60	2/60	2/60	2/60	2/60	2/60

Subject13	2/120	2/120	2/120	2/120	2/120	2/120	2/120
Subject14	4/120	4/120	4/120	4/120	4/120	4/120	4/120
Subject15	0/120	0/120	0/120	0/120	0/120	0/120	0/120
Subject16	4/120	4/120	4/120	4/120	4/120	4/120	4/120
Subject17	0/120	0/120	0/120	0/120	0/120	0/120	0/120
Subject18	2/120	2/120	2/120	2/120	2/120	2/120	2/120
Score	B						

Figure 8.28c: Schedule of a two-factor repeated-measures design with repeated measures on the first factor, with the nine level combinations from Figure 8.23, for the third week, with pretests (B)

Day	22	23	24	25	26	27	28
Subject1							
Subject2							
Subject3							
Subject4							
Subject5							
Subject6							
Subject7							
Subject8							
Subject9							
Subject10							
Subject11							
Subject12							
Subject13							
Subject14							
Subject15							
Subject16							
Subject17							
Subject18							
Score	A						

Figure 8.28d: Schedule of a two-factor repeated-measures design with repeated measures on the first factor, with the nine level combinations from Figure 8.23, for the fourth week, with posttests (A)

Day	29	30	31	32	33	34	35	36
Subject1	4/0	4/0	4/0	4/0	4/0	4/0	4/0	
Subject2	2/0	2/0	2/0	2/0	2/0	2/0	2/0	
Subject3	4/0	4/0	4/0	4/0	4/0	4/0	4/0	
Subject4	0/0	0/0	0/0	0/0	0/0	0/0	0/0	
Subject5	2/0	2/0	2/0	2/0	2/0	2/0	2/0	
Subject6	0/0	0/0	0/0	0/0	0/0	0/0	0/0	
Subject7	4/60	4/60	4/60	4/60	4/60	4/60	4/60	
Subject8	2/60	2/60	2/60	2/60	2/60	2/60	2/60	
Subject9	4/60	4/60	4/60	4/60	4/60	4/60	4/60	

Subject10	0/60	0/60	0/60	0/60	0/60	0/60	0/60	
Subject11	2/60	2/60	2/60	2/60	2/60	2/60	2/60	
Subject12	0/60	0/60	0/60	0/60	0/60	0/60	0/60	
Subject13	4/120	4/120	4/120	4/120	4/120	4/120	4/120	
Subject14	2/120	2/120	2/120	2/120	2/120	2/120	2/120	
Subject15	4/120	4/120	4/120	4/120	4/120	4/120	4/120	
Subject16	0/120	0/120	0/120	0/120	0/120	0/120	0/120	
Subject17	2/120	2/120	2/120	2/120	2/120	2/120	2/120	
Subject18	0/120	0/120	0/120	0/120	0/120	0/120	0/120	
Score	B							A

Figure 8.28e: Schedule of a two-factor repeated-measures design with repeated measures on the first factor, with the nine level combinations from Figure 8.23, for the fifth week, with pretests (B) and posttests (A)

Answers to Chapter 9

9.1 a. A single-case study should be used if one intends to prove that a treatment has an effect for one particular subject, which differs essentially from all comparable subjects, e.g., because of the individual history.

b. Single-case studies should be used if treatments are compared with respect to their effectiveness and if one cannot rule out that the effects of the treatments are not the same for different subjects though the reasons for these differences may not be known.

9.2 a. In an intervention study it can only then be concluded that a treatment has had an effect if the time of intervention is randomly selected under many similar points of time, and if a true double-blind study is performed, i.e. neither the subject nor the experimenter knows the time of intervention and both have no means to detect it.

b. In a comparison of different treatments, it can only then be concluded that treatment differences exist, if the timely order of treatments is randomly fixed and neither the subject nor the experimenter can find out which treatment is effective at a given point of time.

9.3 If the conditions for Edgington experiments (i.e., randomization of times of intervention or of treatments, true double-blind studies) are fulfilled, causal conclusions can be drawn, because at least in the statistical sense, all possible alternative explanations can be ruled out.

9.4 a. If due to the randomization procedure designs are obtained in Edgington experiments which correspond to designs which without a randomization would admit obvious alternative explanations, we have the problem that it may be suspected that a detected effect might be caused by a statistical Type I error.

b. Any doubt that the double-blind condition has really been established weakens the interpretability of Edgington experiments.

9.5 For 11 cups with 5 M-cups and 6 T-cups result
$(11 \times 10 \times 9 \times 8 \times 7) / (5 \times 4 \times 3 \times 2 \times 1) = 462$
possible arrangements.

9.6 Because, as a rule, it is not required for Edgington experiments that the subjects are asked to identify the design which was assigned to them and that this answer is used as a dependent variable, as this was the case for the tea-tasting experiment, no open protocol is needed. The informed consent of the participants of a study which may be necessary for ethic or legal reasons may even become a threat for the double-blind condition.

9.7 If a factor has the three levels C (control), T1 (first treatment), and T2 (second treatment) and if each level is to appear four times, there (with $n! = 1 \times 2 \times ... \times n$) exist

$(3 \times 4)! \,/\, (4! \times 4! \times 4!) = 34650$
possible different arrangements of which one has to be randomly selected. For this, four cards with "C", four cards with "T1", and four cards with "T2" are laid into a box, the cards are carefully shuffled, and then one card after the other is drawn without replacement from the box. This fixes the arrangement of the 12 experimental conditions.

9.8 If the duration of the intervention is to be, e.g., 5 days and the duration of the study 130 days, the first intervention day is randomly selected from the days 11 to 110, where the intervention takes place not only on the selected day but also on the 4 following days. As a measure for a treatment effect, the sum of the scores at the five intervention days is used. If delayed effects of the intervention are suspected, it is also possible, e.g., to use the sum of the scores at the 5 intervention days and at the 5 succeeding days as a measure of an intervention effect.

An even more efficient design would be obtained if the number of intervention days is not kept fixed but may also vary randomly between 1 and 10 days. As a measure of an intervention effect the sum of the scores of all intervention days may be used. The first intervention day is randomly selected from the days 11 to 110 as above. After that an integer is randomly selected from 1 to 10. If a 1 is selected, only one intervention day is considered. If a 2 is selected, a second intervention day is considered in addition, etc. With 100 possible ways to select the first intervention day and with 10 possible ways to select the number of intervention days, we obtain 100 × 10 = 1000 different designs, all with the same probability (.001). In contrast, only 100 designs were obtained for the simple intervention design above.

9.9 The sum of the P-values is given by

$$S = .1 + .3 + .01 + .6 + .65 + .2 + .01 + .03 + .1 + .04 + .02 + .01 = 2.07,$$

and this yields

$$P_{\mathrm{T}} = \frac{2.07^{12}}{12!} - \frac{1.07^{12}}{1!11!} + \frac{.07^{12}}{2!10!} = .000013.$$

The complementary P-values are given by

.9, .7, .99, .4, .35, .8, .99, .97, .9, .96, .98, .99

and therefore their sum by

$$S^C = 12 - S = 9.93.$$

For P_T^C we obtain

$$P_T^C = \frac{9.93^{12}}{12!} - \frac{8.93^{12}}{1!11!} + \frac{7.93^{12}}{2!10!} - \frac{6.93^{12}}{3!9!} + \frac{5.93^{12}}{4!8!} - \frac{4.93^{12}}{5!7!} + \frac{3.93^{12}}{6!6!} - \frac{2.93^{12}}{7!5!} + \frac{1.93^{12}}{8!4!}$$

$$-\frac{.93^{12}}{9!3!} = .999987$$

and therefore

$$P_T = 1 - P_T^C = .000013.$$

The given value for P_T^C is obtained only then if the single terms in P_T^C are computed with a sufficient number of digits, e.g., with 8 digits following the decimal point.

9.10 a. In view of the large number of statistical comparisons which were performed, even under optimal conditions there is a high probability that the existence of non-existing effects is proved (cf. Section 3.1.3), i.e. the pretended results may be due to random inhomogeneity of the data.

b. Because the more than 24 measurements for each subject must be considered as dependent, the assumption of independence which is fundamental for an analysis of variance is just as questionable as the assumptions of normal distributions and homogeneity of variances (cf. Section 3.1.2 and 7.3).

c. With respect to the factor "drug" a simple crossover design was used. For this it is not admissable, to pool the drug and placebo samples from day 1 and day 2 to achieve in this way an artificial doubling of the sample sizes (cf. Section 6.4).

d. The 6 experimental conditions which arise from the factor "drug" with 2 levels and the factor "position" with 3 levels, were not randomly assigned to the subjects. In this case, it would have been possible to assign randomly to each subject one of the 6! = 720 possible designs and to combine afterwards the outcomes resulting from the single subjects (cf. Section 9.4). Because it might have been difficult to realize this proceeding in practice, it would have been alternatively possible to fix for each subject the order of drug and placebo in a random way, in contrast to the study, where only counterbalancing was used. After this, within each day the order of positions had to be randomized, as it is also described in the study. In this way, $3! \times 2 = 6 \times 2 = 12$ different designs would have resulted which could have been randomly assigned to the subjects. Because of $(1 / 12) = .083$, for no subject a statistically significant result could have been expected. However, by combining the P-values over the subjects a significant result might have been found.

It cannot be ruled out, however, that the discussed ways to replace the design of the study by single-case designs, probably would also not yield interpretable outcomes. It is not very likely that the assumption that the subjects were blind with respect to the position of the rectal balloon, is justified. Even if the subjects could not describe the position of the balloon in a correct way, and this, it seems, was not investigated, it cannot be ruled out that the subjects could differentiate between the three positions. This might result in biased outcomes due to hypothesis guessing (cf. Section 3.3.4) and social desirability responding (cf. Section 3.3.5). Further, experimenter effects (cf. Section 3.3.6) cannot be ruled out, because the experimenter was not blind with respect to the position of the balloon, as it is also established in the study.

Therefore, for the investigated problems, by no kind of multiple-treatment designs, and this is true also for single-case experiments, we could have obtained outcomes which allow a causal interpretation. Rather the original sample of 18 subjects should have been randomly split up into 6 subsamples with 3 subjects each, which would have been randomly assigned to the 6 experimental conditions. Because the sample size for each condition most probably would have been too small to prove the existence of effects by means of statistical tests, it would have been better to start with 36 or even 54 subjects. This would have been more acceptable for ethic as well as for economic reasons than the actually performed study with 18 subjects for which it was obvious from the beginning that it could not yield outcomes with a reasonable interpretation.

If an increase of the sample size would not be feasible for some reasons or others and if with only three subjects for each subsample a detection of effects could not be expected, there is left the possibility to split up the original sample at random into two subsamples with nine subjects each, to be at least able to detect an effect of the factor "drug" if such an effect should exist. Alternatively, the original sample could be randomly split up into three subsamples with six subjects each, to detect an effect of the factor "position" if it exists.

9.11 a. On the basis of the data of only 20 subjects 7 dependent statistical tests are performed. Therefore, it cannot be ruled out that the pretended effects are only artifacts due to random heterogeneity of the data (cf. Section 3.1.3).

b. Because the arrangement of the 8 conditions was fixed for each subject at random, it might seem at first glance that for each subject a single-case experiment was performed, where one of 8! = 40320 possible designs was randomly assigned to each subject. However, this is not true: after fixing the time of day and the radio condition the two task difficulty conditions followed one after another, though in a randomized way. There are 4! = 24 ways to order the 4 time of day-radio combinations and for each of these arrangements there are 2 ways to order the levels of task difficulty. Therefore, not 8! but only 24 × 2 = 48 possible designs arise. After performing a distribution-free statistical evaluation (cf. Section 9.1) it would have been possible to combine the results for all 20 subjects (cf. Section 9.4).

c. However, the evaluation performed in the study as well as the proceeding proposed above will not yield results which permit a reasonable interpretation because the subjects (just as the experimenter) obviously were not blind with respect to the 8 experimental conditions and had no difficulty to discriminate them. Therefore, biases due to hypothesis guessing (cf. Section 3.3.4), social desirability responding (cf. Section 3.3.5), and experimenter effects (cf. Section 3.3.6) cannot be ruled out.

Instead of a multitreatment design which due to the absent blinding does not yield interpretable outcomes, an original sample, of 64 subjects say, should be randomly split up into 8 subsamples of size 8 each, and these subsamples should be assigned to the 8 factor level combinations in Figure 9.3. After the respective treatment, for each subject only one measurement of the dependent variable (here: accumulated deviations of the car position from the midline of the road) would be recorded.

1. MRE	5. ARE
2. MRD	6. ARD
3.MNE	7. ANE
4. MND	8. AND

Figure 9.3: Three-factor design with the factors time of day (M = morning, A = afternoon), radio (R = radio, N = no radio), and difficulty (E = easy, D = difficult)

If the effects of the simultaneous influence of the three factors are of no interest, it is possible to restrict oneself to the three one-factor designs with two levels each, which are depicted in Figure 9.4. For this a sample of, say 48 subjects, is randomly split up into 6 subsamples with 8 subjects each, and to each of the 6 conditions in Figure 9.4 one of these subsamples is randomly assigned. For the first one-factor design, e.g., the conditions R and E could be kept constant, for the second design the conditions M and E, and for the third design the conditions M and R.

M	A

R	N

E	D

Figure 9.4: One-factor designs for the factors time of day (M = morning, A = afternoon), radio (R = radio, N = no radio), and difficulty (E = easy, D = difficult)

References to the Questions

[1] Bongard, S., Hodapp, V., Frisch, M., Lennartz, K. (1994). Effects of active and passive coping on task performance and cardiovascular reactivity. Journal of Psychophysiology 8, 219-230.

[2] Crowell, M. D., Cheskin, L. J., Musial, F. (1994). Prevalence of gastrointestinal symptoms in obese and normal weight binge eaters. American Journal of Gastroenterology 89, 387-391.

[3] Crowell, M. D., Musial, F., French, A. W. (1993). Eating lowers defecation threshold in pigs through cholinergic pathways. Physiology & Behavior 53, 1029-1032.

[4] Hehl, F. J., Ruch, W. (1990). Conservatism as a predictor of responses to humour-III. The prediction of appreciation of incongruity-resolution based humour by content saturated attitude scales in five samples. Personality and individual Differences 11, 439-445.

[5] Jäncke, L., Musial, F., Vogt, J., Kalveram, K. T. (1994). Monitoring radio programs and time of day affect simulated car-driving performance. Perceptual and Motor Skills 79, 484-486.

[6] Kröner-Herwig, B., Diergarten D., Diergarten, D., Seeger-Siewert, R. (1988). Psychophysiological reactivity of migraine sufferers in conditions of stress and relaxation. Journal of Psychosomatic Research 32, 483-492.

[7] Kröner-Herwig, B., Fritsche, G., Brauer, H. (1993). The physiological stress response and the role of cognitive coping in migraine patients and non-headache controls. Journal of Psychosomatic Research 37, 467-480.

[8] Musial, F., Crowell, M. D., Kalveram, K. T., Enck, P. (1994). Nutrient ingestion increases rectal sensitivity in humans. Physiology & Behavior 55, 953-956.

[9] Musial, F., Enck, P., Kalveram, K. T., Erckenbrecht, J. F. (1992). The effect of loperamide on anorectal function in normal healthy men. Journal of Clinical Gastroenterology 15, 321-324.

PART B

Dictionary of Experimental Design

ABAB design: Extension of the **ABA design** by an additional treatment phase (B). No **causal conclusions** are possible.

ABA design: Special **within-subjects design**, where first a **baseline** (A) is recorded, then a treatment (B) is applied, and then again a baseline (A) is recorded with the same subjects. Because of **maturation**, **history**, and other plausible **alternative explanations** this design and its extensions do not permit **causal conclusions**. By a comparison of the two baselines one tries to find out whether the treatment effect is reversible. The term **withdrawal design** is also used.

ABCB design: Corresponds to an **ABAB design**, where the second **baseline** phase is replaced by a **placebo** phase C. **Causal conclusions** are not possible.

AB design: A **baseline** phase (A) with one or more measurements is followed by a treatment phase (B) with one or more measurements. **Causal conclusions** are not possible.

Abrupt change: See **delayed causation**.

Absolute effect: If a **treatment group** is compared to a **control group** by means of an **independent two group design**, where no treatment or only a sham treatment is applied to the control group, a difference of the two groups with respect to a **dependent variable** is called an absolute effect in contrast to a **relative effect**.

Acceptance sampling: A **sample** is drawn from a **population**. This sample is accepted if the proportion of not suitable **sample units** with respect to a given standard of quality does not exceed an a priori fixed percentage within the sample.

Accidental sample: A **sample** from a **population** which is just available, e.g., students in a lecture or customers in a warehouse. Due to the threat of **selection effects** the population for which such a sample might be representative is not known.

Accrual rate: In **clinical trials** of long duration the accrual rate, i.e., the number of patients which enter the study per time unit, should be controlled. Then it is possible to intervene if the accrual rate falls below a fixed rate. In particular, in **multicentre trials** an approximately equal accrual rate should be striven for, in order to avoid differing numbers of patients.

Accuracy: The accuracy is high, if the **precision** is high and if no systematic errors are present. A low accuracy can result due to systematic errors even if the precision is high.

Acquiescence: See **response bias**.

Acquiescence response style: See **acquiescence** and **yeasaying**.

Active control equivalence study: A study by which one wants to prove that a new treatment has the same **efficacy** as a standard treatment. Such a study might be advisable if **adverse effects** have to be expected for a standard treatment for all patients or for a part of them. As it is not possible, in principle, to prove that both treatments have exactly the same effect, one can only try to show that any differences will not exceed a given amount.

Active control trial: Clinical trial, where a **relative effect** is to be shown by the comparison of a new drug with another drug. Here, no **absolute effect** is to be demonstrated by comparing the drug with a **placebo**.

Active variable: Independent variable which is controlled by the researcher.

Adaptation phase: In particular, in case of physiological measurements a certain **reactivity** has to be expected. Therefore, an adaptation phase should follow after each treatment phase, until a **baseline** has been established which is similar to the baseline before the treatment phase. This might not be possible if **irreversible effects** of a treatment occur and in such cases an adaptation phase can only serve to obtain a stable behavior.

Adaptive cluster sampling: Subjects are selected from a **population** for an initial sample, and measurements are recorded for these subjects. Always if these measurements obey certain **inclusion criteria** for a given subject, additional subjects from a suitably defined neighborhood of the subject are included into a **sample**.

Adaptive design: A design for a clinical trial, in which the selection of a treatment for a patient depends on the effects which this treatment had on patients previously treated in the study. Examples are the **play-the-winner rule** and the **randomized play-the-winner rule**.

Adaptive sampling: In case of a successive selection of **sample units** for a **sample** from a

population, the selection of further sample units is made dependent on the values which certain variables have exhibited for sample units which were selected up to the present moment.

Adherence: Synonym for **compliance**.

Ad hoc sample: Synonym for **accidental sample.**

Adjugate Latin square: An adjugate Latin square can be derived for a given **Latin square** by exchanging the treatment number and the column number. In the following example the second Latin square is an adjugate Latin square with respect to the first one. E.g., we find the fourth treatment (D) in the first Latin square in the second row in the first column. Therefore, we find the first treatment (A) in the adjugate Latin square in the second row in the fourth column. Another adjugate Latin square would result, if the treatment number is interchanged with the row number.

A	D	C	B
D	A	B	C
C	B	D	A
B	C	A	D

A	D	C	B
B	C	D	A
D	B	A	C
C	A	B	D

Adverse effect: Undesired **side-effect**.

Age effect: An **effect** which is caused by the age of a subject. In contrast, **period effects** are caused by the fact that subjects live in certain historical epochs, while **cohort effects** are due to the fact that each subject forms part of a particular generation.

Age heaping: This occurs if the precise age of subjects is not recorded but if the age is only given in rough terms, e.g., only the year but not the day of birth.

Age-period-cohort model: Often it is reasonable to assume that the risk of a disease depends on the age of the subjects, on the age of the subjects at their first contact with the **risk factor**, and on the duration of the contact with the risk factor. The sum of the last two times yields the first one.

Aggregate: A set of **sample units**, which have the same levels with respect to one or more variables, e.g., the inhabitants of a town or the patients of a hospital.

Aggregate data: Data from **aggregates** contain information about aggregates but not about the single **sample units** of which an aggregate is composed. Data about single sample units are also called **micro-data**.

Alias: See **confounded factorial effect**.

Aligned systematic sampling: A geographic region is subdivided into quadrats of the same size. In one quadrat a subject is selected at random. Then, from each of the other quadrats a subject is selected which is located within the quadrat at the same place as the subject which was selected from the first quadrat.

Alternative explanation: In a study which is not well planned, there exist one or more alternative explanations for any observed relation between an **independent** and a **dependent variable** apart from the preferred explanation.

Analysis as-randomized: Synonym for **intention-to-treat analysis**.

Analytical experiment: According to Robinson (1976, p. 173) this is an experiment, where samples of subjects are assigned to several **levels** of an **independent variable** by means of a randomization. By such an experiment it is possible, in contrast to an **exploratory experiment**, to find out which levels of an independent variable cause which values of a **dependent variable**.

Anecdotal evidence: Information which is not the result of **causal conclusions** based on the outcome of an experiment but which results from random observations.

Animal model: Experiment with animals instead of human beings, where a certain transferability of the results for the animals to men is assumed.

Antecedent variable: Variable which occurs in a **causal chain** before another variable.

Aptitude-treatment interaction: Trait-treatment interaction where aptitude is considered as a trait.

Area sampling: A geographical region is divided into precisely defined sub-areas from which a **sample** is selected. From these selected sub-areas all subjects, subsamples of subjects or single subjects are selected.

Arms of a study: If, as, e.g., for an **intervention study**, an initial sample of subjects is subdivided over and over again in the course of the study, where the subsamples are exposed to different experimental conditions, the endpoints of this ramification are also called the arms of the study. Therefore, in an arm of a study a subsample of subjects is found, all of which have experienced the same sequence of experimental conditions in a fixed order.

Artifact: Non-existing effect which is wrongly identified because possible **alternative explanations** are not taken into account.

Artificial pairing: Forming pairs of subjects by means of **matching variables** when **matching** is being used.

Artificial selection: See **selection**.

Ascertainment bias: In particular, in **retrospective studies** one cannot rule out that certain levels of a variable are observed far more often in a sample of patients than in healthy subjects solely for the reason that the expectancies of the researcher result in a selective attention with the consequence that an equal rate of occurrence of the same levels of the variables in healthy subjects is overlooked.

Assay: An experiment by which the strength or nature of the effect of a **causal variable** on the response of a subject is investigated. Also see the key-word **bioassay**.

Assay run: Sequence of measurements at the same sample of subjects.

Assigned treatment: Treatment which is assigned to a patient after he or she has entered a **clinical trial**.

Asymmetric carry-over effect: If a sequence of two or more experimental conditions is applied to a subject, one cannot rule out that asymmetric carry-over effects occur, i.e., **carry-over effects** which change if the timely order of the experimental conditions is altered. If such asymmetric carry-over effects occur in **within-subjects designs**, often the effects of the single experimental conditions cannot be isolated and, therefore, no **causal conclusions** with respect to these conditions are possible.

Asymmetric transfer: Synonym for **asymmetric carry-over effect**.

Attrition: Loss of subjects in a **longitudinal study** due to **experimental mortality**.

Attrition bias: One can never rule out that the measurements for **dropouts**, if it had been possible to record them, differ considerably from the measurements for the subjects which remained in the study.

Attrition rate: Rate of **dropouts** with respect to the original **sample size** .

Autochthonous variability: Variability which is caused by **endogenous variables**.

Available sample: In case of **samples** which are used in real studies it is often unknown for which **populations** they are **representative samples**, because they are, e.g., neither **random samples** nor **quota samples**, but just samples of subjects which are available for a researcher who wants to perform a study. Therefore, it is not known to which populations any **causal conclusions** can be generalized.

BAB design: A treatment phase (B) with several measurements is followed by a **baseline** phase (A) with several measurements which in turn is succeeded by a treatment phase (B) with several measurements. **Causal conclusions** are not possible.

Background variable: Causal variable which cannot be manipulated, as, e.g., gender.

Backward-looking study: Synonym for **retrospective study**.

Balaam's design: An extension of a **crossover design** which is sometimes named after Balaam (1968). By using the same experimental conditions more than once it becomes possible to investigate also **carry-over effects**. In particular, for two experimental conditions A and B the simple crossover design with the **sequences** A_1B_2 and B_1A_2 can be extended to a Balaam design with the sequences A_1A_2, A_1B_2, B_1A_2, and B_1B_2.

Balanced change-over design: Synonym for **balanced crossover design**.

Balanced crossover design: Crossover design, where each sub-sequence of two treatments occurs equally often. E.g., each of the four sub-sequences A_1B_2, B_1A_2, A_1A_2, and B_1B_2 occurs exactly three times in a design with the four **sequences** $A_1B_2B_3A_4$, $B_1A_2A_3B_4$, $A_1A_2B_3B_4$, and $B_1B_2A_3A_4$.

Balanced design: In the most simple case an **experimental design**, in which the same number of subjects is assigned to each combination of factor levels.

Balanced incomplete block design: A design with b **blocks** each with k subjects is considered, where each of t treatments is applied exactly r times. In a balanced incomplete block design each pair of treatments occurs exactly c times in the same block. From this follow the two relations $bk = rt$ and $c(t - 1) = r(k - 1)$. In the following example we have $t = 4$, $k = 2$, $r = 3$, $b = 6$, and $c = 1$. Here, each of the 6 cells corresponds to a block with 2 subjects, and A, B, C, and D correspond to 4 treatments, each of which occurs 3 times. E.g., the pair AB occurs only in $c = 1$ block.

AB	AC	AD
CD	BD	BC

Balanced Latin square: A balanced Latin square is a **Latin square**, in which each treatment occurs in the rows once before and once after each other treatment. A construction like this is only possible if the number of rows is a square number. It should be noticed in the following example that the required property holds only for the rows.

A	B	D	C
B	C	A	D
C	D	B	A
D	A	C	B

Balanced lattice square: See **lattice design**.

Balanced longitudinal data: Data from a **longitudinal study**, for which the time intervals between comparable measurements have the same length for all subjects, though the intervals between adjacent measurements for a single subject do not have to have the same length. Further, for each subject measurements have to be available for all considered points of time.

Balanced paired comparison design: Synonym for **round robin design**.

Balanced sample: A **sample** which is selected such that the mean with respect to a given **manifest variable** is the same as the corresponding mean in the **population**, is called balanced sample. Of course, it does not necessarily hold that this property of the balanced sample holds also for other manifest variables.

Balancing: Balancing is a technique for the **control of extraneous variables**. Here, one tries to assign levels or level combinations of supposed **extraneous variables** in a systematic way uniformly to the different levels or level combinations of the **independent variables**. The concept of balancing should not be mixed up with the concept of **counterbalancing**.

Baseline: In a **time-series design** a sequence of measurements is recorded before applying a treatment. These measurements should give information about which variations of the scores or which trends in the data can be expected even without any treatment. Such pretest scores which all together constitute the baseline are recorded up to that point of time, where a stable, i.e. a predictable, state is achieved. Here, each researcher has the problem to exactly define the meaning of "stability" of a baseline.

Baseline balance: Non-existence of important differences between **baselines** for different groups of subjects.

Baseline characteristics: Measurements at subjects before they are exposed to different experimental conditions.

Basement effect: Synonym for **floor effect**.

BBD: Abbreviation for **binary block design**.

BCOD: Abbreviation for **balanced crossover design**.

B design: Only one treatment phase (B) with several measurements during the treatment phase is considered. **Causal conclusions** are not possible.

Before-after design: A **dependent variable** is measured for a sample of subjects. Then, a treatment is applied. After this, the dependent variable is measured a second time. In most cases this does not differ from the **one-group before-after design**. Because no **control group** is used, **causal conclusions** are not possible.

Before-after static group comparison design: According to Matheson et al. (1971, pp. 43-44) this is a **static group comparison design**, where the **dependent variables** are recorded before and after the experimental conditions are effective. Due to the absence of a **randomization**, no **causal conclusions** are possible.

Before-after two group design: According to Matheson et al. (1971, pp. 45-46) this is an **independent two group design**, where the **dependent variables** are recorded before and after the experimental conditions are effective. Though it is possible, in principle, due to the presence of a **randomization** to draw **causal conclusions**, it might be difficult to interpret the outcome due to the possible **reactivity** of the pretest measurements.

Before-match-after design: According to Matheson et al. (1971, pp. 48-49) the only difference with respect to the **match by correlated criterion design** is that the **matching variable** is equal to the **dependent variable**. If a **reactivity** of the pretest measurement cannot be ruled out, **causal conclusions** might not be possible. A pretest might have, e.g., in a **treatment group**, another effect as in a **control group**.

Behavioral unit: Synonym for **experimental unit** or **sample unit**.

Bellwether precinct: District in which a **Bellwether sample** is selected.

Bellwether sample: According to Matheson et al. (1971, p. 27) this is a small systematically selected sample, e.g., inhabitants of a certain town, which are known to show, e.g., with respect to an election, the same behavior as the total **population**, e.g., the people of the country, where the town is located.

Benchmark: See **benchmarking**.

Benchmarking: A method used in order to bring less reliable measurements in accord with more reliable measurements which are called **benchmarks**.

Berenblut design: Certain **crossover designs** which permit the study of simple **carry-over effects** and which were considered among others by Berenblut (1964). For the two experimental conditions A and B the following four **sequences** $A_1B_2B_3A_4$, $B_1A_2A_3B_4$, $A_1A_2B_3B_4$, and $B_1B_2A_3A_4$ are used, each with four **periods**. In a Berenblut design the experimental conditions are assigned to each sequence and to each period the same number of times (two times in the example). Further, each sub-sequence of two different or equal conditions occurs exactly once in each sequence of two succeeding periods, and each sub-sequence of two different or equal conditions occurs equally often if all sequences of two succeeding periods are considered (three times in the example).

Berkson's fallacy: Illusory correlation between the occurrence of two diseases or between a disease and a supposed **risk factor**, which occurs because the composition of the group of patients which enter a **clinical trial** deviates considerably from the composition of the group of patients in the corresponding **population** (Berkson, 1946).

Between-groups design: Synonym for **between-subjects design**.

Between-subjects design: In contrast to **within-subjects designs**, in between-subjects designs to each experimental condition or to each combination of such conditions, respectively, another sample of subjects is assigned. For between-subjects designs, in general, **causal conclusions** are possible, if a **randomization** took place.

Between-subjects factor: Synonym for **between-subjects variable**.

Between-subjects variable: Independent variable, for which for each subject exactly one **level** is considered.

Bias: Anything that makes **alternative explanations** possible for outcomes of studies.

Biased coin method: If patients enter a study sequentially and have to be assigned to a **treatment** or **control group** immediately, various procedures can be used if a **balancing** with respect to a **block factor** is intended in addition. See, e.g., the key-word **permuted blocks within strata**. Efron (1971) proposed the biased coin method. For this, first a probability p with $.5 < p < 1$ is fixed, e.g., $p = 2 / 3$. If a patient enters a study and is assigned to a certain block, three situations are possible: 1. Up to now either no patients belong to the block or an equal number of patients for both conditions. In this case the patient is assigned to the treatment group with probability .5. For this a fair die can be used, where for the outcomes 4, 5 or 6 the patient is assigned to the treatment group, and for the outcomes 1, 2 or 3 to the control group. 2. Within the block a majority of patients have already been assigned to the treatment group. In this case the new patient is assigned with probability p to the control group or, respectively, with probability $(1 - p)$ to the treatment group. E.g., with $p = 2 / 3$, the new patient would be assigned to the control group in case of the outcomes 3, 4, 5 or 6, but in case of the outcomes 1 or 2 to the treatment group. 3. Within the block a majority of patients have already been assigned to the control group. In this case, the new patient is assigned with probability p to the treatment or, respectively, with probability (1 – p) to the control group. For more than two experimental conditions, a generalization of the biased coin method was proposed by Pocock (1979).

Biased sample: A **sample** that deviates in its composition from the **population** from which it was drawn.

BIBD: Abbreviation for **balanced incomplete block design**.

Binary block design: A **block design** in which each experimental condition either does not occur in a given **block** or occurs exactly once.

Bioassay: An experimental method used in order to study the effectiveness of a drug in organisms. In a **direct assay** the dose of the drug under study is increased up to that point of time, where a particular response is observed which was fixed in advance. The corresponding maximum dose is also called **tolerance level**. In an **indirect assay** several fixed doses of a drug are applied to organisms and the extent of a response is measured. The found relation between the size of the dose and the extent of the response is called **dose-response curve**. If such a relation is described by a linear function, we have a **slope ratio assay**. If the extent of the response is described by a linear function of the logarithm of the dose, we have a **parallel line assay**.

Bioavailability study: Study of those variables which influence the size of that portion of an applied dose of a drug which reaches the location where it is effective and of the velocity with which this location is reached.

Bioequivalence trial: One investigates whether two different compounds which both contain the same effective agent have the same effect. Since it is not possible, in principle, to prove the equality of effects, one tries to show that a possibly existing difference does not exceed a fixed bound.

Biological assay: Synonym for **bioassay**.

Biological efficacy: Biological effect of a treatment for all subjects to whom the treatment, to which they were assigned, was applied.

Birth-cohort study: Synonym for **cohort study**, if the **cohort** consists of subjects of the same age.

Blank experiment: Usage of irrelevant experimental conditions in a nonsystematic way to avoid automatic responses of subjects.

Blind analysis: Synonym for **blinded evaluation**.

Blinded evaluation: See **blinding**.

Blindfold experiment: See **blinding**.

Blinding: In **single-blind studies** or **single blindfold experiments** a subject does not know which experimental condition it has been assigned to. In **double-blind studies** or **double blindfold experiments** neither the subject nor the experimenter do know which experimental condition is present. In **triple-blind studies** or **triple blindfold experiments** neither the subject nor the experimenter nor the subject who evaluates the study (**blinded evaluation**) do know which experimental condition is present.

Block: See **blocking**.

Block design: One **extraneous variable** is considered and one **block** is assigned to each of its levels.

Block factor: Synonym for **block variable**.

Blocking: A **local control** technique, where subsamples of subjects are formed such that given known **extraneous variables** have the same levels for all subjects of a subsample, i.e. are kept constant for this subsample. These subsamples are called **blocks**. The considered extraneous variables are called **block variables**. Instead of blocking the term **matching** is used, if each block consists of only two subjects. Each block contains as many subjects as there are experimental conditions. The experimental conditions are assigned to the subjects of a block by means of a **randomization**.

Blocking off: This is a **global control** technique where the subjects are blocked off against distracting stimuli.

Block randomization: Synonym for **permuted block design**.

Block variable: See **blocking**.

Booster treatment: A reapplication of a treatment in a **follow-up study**.

Border effect: Border effects can occur if the **level** of an **independent variable** does not only affect subjects which are assigned to this level but also subjects which are neighbors in space or time to the first subjects, though they are assigned to other levels.

Borrowing effect: In cases, where the relative frequencies add to one, very small or very large frequencies, respectively, of an event with respect to another **population** might be due to a borrowing effect. This consists in the observation that a decrease of one relative frequency necessarily must cause the increase of other relative frequencies and vice versa. If, e.g., the different causes of death in a subpopulation are being considered, and one finds out that a certain cause has a low relative frequency in comparison with the total population, it might well be that the relative proportion of subjects dying due to this cause is the same in both populations, but that other causes have affected a higher percentage of deaths in the subpopulation than in the total population. Assume, e.g., a population of 1000 subjects, where 100 subjects die due to cause A and another 100 subjects due to cause B. Hence, fifty percent of the deaths are due to cause A in the total population. Now, consider a subpopulation of 100 subjects, where 10 subjects die due to cause A and 30 subjects due to cause B. In this subpopulation, only twenty-five percent of the deaths are due to cause A, though in both populations ten percent die due to this cause.

Buffer: Irrelevant tasks which are intermingled with the relevant tasks in a **within-subjects design** in the hope to get independent responses to the relevant tasks.

Bulk sampling: Selection of a **sample** from an available group of subjects.

Calendarization: One tries to convert dates which have been recorded with respect to different time units, e.g., months and years, into the same units to get, on the whole, more reliable measurements.

Caliper matching: In a **match by correlated criterion design** subjects are paired by means of a **matching variable**. If limits are fixed for the matching variable which define the range in which the difference of the values for a pair might vary, this is called caliper matching.

Camera silens: A room that is isolated from the outside with respect to acoustic and optic stimuli.

Capture-recapture sampling: The subjects of a **population** either exhibit a certain characteristic, e.g., a disease, or they do not. An estimate of the total number of subjects exhibiting the characteristic without performing a **census** is being sought. For this, first a **random sample** is selected from the population and it is established which subjects in the **sample** exhibit the characteristic. After this, a second sample is selected from the population which is totally independent of the first random sample. Then, one finds out the number of subjects which exhibit the characteristic and who were found in the first sample and who are also found in the second sample.

Carry-over effect: If several experimental conditions are applied to the same subject in a timely order, one cannot rule out that the value of a **dependent variable** is influenced not only by the directly preceding experimental condition but also by more remote conditions. As a consequence, it is often not possible to isolate the effects of single experimental conditions in **within-subjects designs**, such that **causal conclusions** with respect to these conditions are not possible.

Case: A subject in an **epidemiological study** which exhibits the investigated disease.

Case-control study: A kind of **retrospective study**, where a group of patients with a certain disease is compared to a group of subjects without this disease with respect to some known **manifest variables**.

Case-crossover design: When a certain event has been observed in a patient, e.g., an attack of asthma, the patient is interviewed with respect to his or her activities and experiences immediately before the event, and in addition one records the number of times the patient is generally exposed to these conditions.

Case-heterogeneity study: In general, the importance of a **risk factor** for a disease is judged by a study on how often the risk factor has been present for patients and healthy subjects. Alternatively or in addition one can investigate, how often the risk factor was present for patients with other diseases, in particular, with such diseases, for which a relation with the risk factor is supposed. This yields alternative control groups in addition to the group of healthy subjects.

Case history method: Synonym for **case study**.

Case study: The behavior of a single subject is observed for a sequence of points of time. No **causal conclusions** are possible.

Causal chain: See **intervening variable**.

Causal conclusion: This is justified, if the conclusion can be drawn that an **independent variable** has an effect on a **dependent variable** and no **alternative explanation** exists.

Causal diagram: Graphical representation of the relations of cause and effect between studied variables by means of arrows.

Causal relation: If a cause has been shown to have led to an effect, a causal relation between cause and effect exists.

Causal variable: Synonym for **independent variable**.

Cause variable: Synonym for **causal variable**.

Ceiling effect: See **floor effect**.

Censored sample: A **sample** of subjects, where the values of the considered **manifest variables** are not known for all subjects. For the missing values it is only known that they are larger than a known threshold, or it is known that they are smaller than a known threshold. See also the key-words **interval-censored data**, **truncated sample**, and **progressively censored data**.

Census: A total **population** is studied, i.e. a census is a **sample** of the size of the considered population and, therefore, by definition a **representative sample**.

Census tract: An exactly defined small geographic region, where a **census** is performed.

Central composite design: Composite design, where a combination of **factor levels** corresponds to a central point in which a maximum or minimum of the **response surface** is expected.

Central composite rotatable design: Symmetric **central composite design**.

Centrally located sample: The proceeding corresponds to that of a **systematic sampling**, solely the number m is not randomly selected from the numbers 1, ..., k, but it is set $m = (k+1) / 2$ for k odd and $m = (k + 2)/2$ for k even.

Change-over design: Synonym for **crossover design**.

Clinical judgement: See **patient withdrawal**.

Clinical method: Synonym for **case study**.

Clinical study: A clinical study is a **prospective study** with patients, where for a given diagnosis the effectiveness of one treatment or the superiority of one treatment over other treatments is to be shown or in which **side-effects** are to be identified.

Clinical trial: Synonym for **clinical study**.

Closed sequential design: See **sequential design**.

Closed sequential sampling: See **sequential sampling**.

Cluster randomization: Instead of assigning single subjects to the experimental conditions using a **random allocation**, whole groups of subjects are randomly assigned to the conditions.

Cluster sampling: If no list of the subjects but only a list of certain groups of subjects (clusters) is available for the selection of a **sample** from a **population**, e.g., a list of hospitals, schools or apartments, a **random sample** is selected from such clusters and all subjects from the selected clusters form the cluster sample. With respect to the subjects, a cluster sample is no random sample.

Code: This is the assignment rule when **blinding** is being used, by which one can identify the subjects to which the experimental or, respectively, the control condition was applied.

COD (*t, p, s*) design: Synonym for **crossover design**.

Cohort: A group of subjects, which all belong to the same class of age. However, the term cohort is often also used for any group of subjects which are observed for a long time interval in a **prospective study**. Therefore, apart from the term cohort, the more restrictive term **birth cohort** is also used, which denotes a group of subjects of the same class of age.

Cohort design: Treatment conditions are assigned to different **cohorts** and their effects are recorded during a certain time interval. Or cohorts are subdivided by means of a **covariate** into **strata** and the change of a **dependent variable** in a time interval is recorded for the different strata (Cook and Campbell, 1979, pp. 126-133). **Causal conclusions** cannot be drawn due to possible **selection effects**.

Cohort effect: See **age effect**.

Cohort study: A **retrospective** or **prospective study**, where one or more **cohorts** are studied for a long time interval with respect to relevant variables.

Combination therapy trial: Clinical trial, where the effect of a combination of treatments is to be investigated. In the most simple case with only two treatments A and B, four groups are required: treatment A and treatment B, treatment A and **placebo**, treatment B and placebo, placebo and placebo.

Combined modality trial: Clinical trial, where the effect of the combination of treatments is studied, which are principally different, e.g., a combination of surgical operations and chemotherapy.

Combined selection: See **selection**.

Community controls: Often, in **observational studies** subjects from the same environment serve as controls, e.g., patients from the same hospital, in order to parallelize origin and environmental factors. Here, the term **hospital controls** is used. As an alternative, subjects might be considered which originate from the same **population** as the target sample. However, this is only possible if this population can be precisely described and if the subjects can be considered to be representative of this population. In such a case the term community controls is used.

Community intervention study: A **clinical trial** where one does not randomly assign single subjects to a treatment but samples of subjects. This is a case of **cluster randomization.**

Comparability: Generic term for **structural equality**, **observational equality**, and sometimes also **representative equality**.

Comparative bioavailability trial: A **bioavailability study** where different kinds of applying a drug are compared with each other with respect to the bioavailability of the drug.

Comparative design: Synonym for **contrast design**.

Comparative trial: Synonym for **controlled trial**.

Comparison group: Often used as a synonym for **control group**. However, as a rule, a comparison group is a control group which has been formed without a proper **randomization**. See also **contrast design**.

Compensatory equalization of treatments: In **field studies** a **treatment group** might receive a treatment which might be considered as extremely desirable from a more general point of view. If now the **control group** gets a kind of compensation, e.g. for political reasons, it might become impossible to detect a treatment effect due to the equalization of conditions.

Compensatory rivalry: A possible effect of a **diffusion of treatments** might be that subjects of the **control group** try to equalize their handicap by increased efforts, whereby the detection of a treatment effect might become impossible. See also **John Henry effect**.

Complementary block: In an **incomplete block design** there exist **blocks** in which not all experimental conditions are present. A block is a complementary block with respect to another block, if just those experimental conditions are present which are missing in the second block but no others.

Complementary block design: In a **block design** each treatment might occur either m_1 times or m_2 times in each **block**. In the corresponding complementary block design a treatment occurs in each block m_2 times if it occurred m_1 times in the original block design and m_1 times if it occurred m_2 times.

Complementary effect: This is, in a **factorial design,** the difference between the common net effect of several **factors** and the sum of the net effects of the single factors. If a first factor A has, e.g., the **levels** A0 (control) and A1 (treatment) and similarly a second factor B the levels B0 (control) and B1 (treatment) and if e(A0B0), e(A0B1), e(A1B0), and e(A1B1) are the effects of the level combinations, the complementary effect is given by (e(A1B1) – e(A0B0)) – ((e(A0B1) – e(A0B0)) + (e(A1B0)

– *e*(A0B0))) = (*e*(A1B1) + *e*(A0B0)) – (*e*(A0B1) + *e*(A1B0)).

Complete balancing: This is present if the same number of subjects is assigned to each combination of **levels** of one or more known **extraneous variables**.

Complete block: A **block**, in which at least one subject is assigned to each of the treatments considered.

Complete block design: A **block design**, in which, in each **block** each of the treatments considered is applied to at least one subject.

Complete counterbalancing: The term denotes a specific form of **counterbalancing** where an equal-sized sample of subjects is randomly assigned to each of the possible arrangements of the experimental conditions. Since **asymmetric carry-over effects** cannot be ruled out, **causal conclusions** might become implausible. In case of three experimental conditions A, B, and C a complete counterbalancing would yield the six arrangements ABC, ACB, BAC, BCA, CAB, and CBA.

Complete cross-classification: See **cross-classification**.

Complete expectancy control: In order to control the **Rosenthal effect**, two **expectancy control groups** are used. In the first group, the existence of an effect is suggested to the experimenter in the **control condition**, in the second group, the non-existence of an effect is suggested to the experimenter in the **treatment condition**.

Complete factorial experiment: A **factorial experiment**, where for each combination of **factor levels** at least one observation is available.

Complete Latin square: A **Latin square**, in which each possible sequence of two treatments occurs the same number of times in the rows and in the columns. In the following example each sequence occurs exactly once.

D	A	C	B
A	B	D	C
C	D	B	A
B	C	A	D

Completely randomized design: An **experimental design**, where the **randomization** which has been used in order to assign treatments to subjects is not restricted. Such a restriction exists, e.g., for a **match by correlated criterion design**, where the treatment can only be randomly assigned within a given pair of subjects.

Complete paired comparison design: A **paired comparison design**, where each pair of subjects is rated at least once.

Complete within-subjects design: According to Underwood and Shaughnessy (1975, p.10, pp. 64-76) this is a **within-subjects design**, where each subject gets all experimental conditions more than once. Here, one tries to achieve that all conditions are equally influenced by **practice effects** or **progressive error**. For this, a suitable systematic timely assignment of the conditions as well as a partial **randomization** within time periods is used. However, other possible effects, such as **reactivity** or **asymmetric carry-over effects** cannot be controlled in an efficient way, such that **causal conclusions** cannot be drawn.

Complex comparison: A comparison of one group with another one is a **simple comparison**. However, if groups are pooled and the arising groups are compared with other groups, these are complex comparisons.

Complex experiment: An experiment which is based on a **factorial design**.

Compliance: The extent, to which patients in a **clinical trial** obey to the imposed rules.

Composite balanced incomplete block design: A **balanced incomplete block design** which preserves this property, if certain **blocks** are omitted.

Composite design: Special **second order design**, where, in a first step, all combinations for two **levels** of a **factor** are used, and where subsequently certain other combinations are considered.

Concomitant factor: Synonym for **covariate**.

Concomitant therapy: Medicaments and other treatments which are applied to participants in a **clinical trial** but which are not related to the trial.

Concomitant variable: Synonym for **covariate**.

Concomitant variation: Occurs, if several variables show a change in the same direction.

Concurrent multiple response design: ABA designs cannot only be extended to more than three time periods but also to the case with two or more **independent variables**. In addition, more than one **dependent variable** can be recorded with the consequence that for each variable a **baseline** of its own must be considered. Even these more complex designs do not permit **causal conclusions**.

Concurrent schedule design: In contrast to a **multiple schedule design**, a subject is not successively but simultaneously exposed to different reference stimuli, which are coupled with different responses to the behavior of the subject.

Conditioning effect: Synonym for **interaction effect**.

Confederate: A subject in a study, which pretends to be a participant though in reality it is an assistant of the experimenter.

Confirmatory experiment: According to McGuigan (1978, p. 75) an experiment by which the knowledge about effects for which much empirical evidence exists, should be made irrefutable. This is the opposite of an **exploratory experiment**.

Confirmatory study: Study which should prove the **efficacy** of a treatment.

Confirmatory trial: Synonym for **confirmatory study**.

Confounded experiment: If the **extraneous variables** are not being controlled (**extraneous variable control**), a **confounding** between independent and extraneous variables can occur. In this case it will no longer be possible to find out whether changes of the **dependent variables** are due to the **independent variables** or to the extraneous variables. This can effect a misinterpretation of the outcome of the experiment. **Causal conclusions** are not possible for a confounded experiment.

Confounded factorial effect: If, in a **factorial design,** the effects of different **factors** cannot be isolated, they will be confounded. Such effects which cannot be split into single effects are also called **aliases**. Confounded factorial effects of factors with **extraneous variables** are also possible.

Confounder: An **extraneous variable**, for which a **confounding** with an **independent variable** is present.

Confounding: A confounding of two **independent variables** is present if a change of one of the independent variables is paralleled by a change of the other one. In such a case it is not clear, which of the independent variables caused a change of the **dependent variable**.

Confounding bias: The bias which is caused by a **confounding**, and which might render a conclusive interpretation of the outcome of a study impossible.

Confounding factor: Synonym for **confounder**.

Confounding variable: Synonym for **confounder**.

Conjugate Latin square: A conjugate Latin square arises from a **Latin square** by interchanging rows and columns, i.e. by a reflection at the main diagonal. Thus, the two following Latin squares can be regarded as conjugate Latin squares.

A	D	B	C
D	C	A	B
C	B	D	A
B	A	C	D

A	D	C	B
D	C	B	A
B	A	D	C
C	B	A	D

Connected block design: A **block design**, in which, for any two treatments, it is possible to form a chain of treatments between these two treatments such that two neighboring treatments are simultaneously applied in the same **block**.

Conservative arrangement of the levels of an extraneous variable: A procedure described by Matheson et al. (1971, p. 24) allowing to take the effects of known **extraneous variables** into account even if the usual control techniques are not being used. In order to do so a profound knowledge about the direction in which the outcome of a study might be biased due to an extraneous variable is necessary. Then, such a level of the extraneous variable is kept constant in the study that an effect of the applied treatment can be found only if it is stronger than a possible effect of the extraneous variable which is effective in the opposite direction. Since no real control of the extraneous variable is exerted, for any found effects **alternative explanations** might exist, i.e. **causal conclusions** are not possible.

Constancy: **Global control** technique, where known **extraneous variables** are kept constant for all subjects.

Constant factor: A **factor** which is kept constant in an **experimental design**, i.e. for which only a single **level** is effective.

Construct: Synonym for **latent variable**.

Construct underrepresentation: According to Cook and Campbell (1979, p. 64) it is possible that, due to a one-sided **operationalization,** not all aspects of a **construct** are reflected in an appropriate way, such that **construct validity** might be threatened.

Construct validity: This is the higher, the better the **manifest variables** which are assigned to the **constructs** by means of an **operationalization** reflect in a correct way the essential properties of the constructs.

Contextual effect: Effect which is caused by the environment of a subject. E.g., the behavior of subjects might be different depending on whether they live in a large town or in the countryside.

Contingency effect: Synonym for **interaction effect**.

Continuous screen design: See **screening study**.

Contrast design: **Quasiexperimental design**, where the **dependent variables** are recorded for observed groups of subjects and where these groups are compared on the basis of these records. Due to the missing **randomization** no **causal conclusions** are possible.

Control condition: Either the experimental condition which is used for a **control group** or that condition in a **within-subjects design** with which a **treatment condition** is to be compared.

Control group: Whether a treatment has an effect on a **dependent variable** cannot be found out by considering a sample of subjects in a **one group before-after design** and by observing, for this **treatment group,** whether the records after the treatment differ from those before the treatment. Any changes are not necessarily due to the treatment but might have been caused, e.g., by **maturation** or **history**. In order to detect the effect of a treatment, the original sample of subjects has to be randomly split (by means of a **randomization**) into a control group and a treatment group. The only difference between the two groups should be that that aspect of the treatment which is to be studied, is present only in the treatment group but not in the control group. A distinct difference of the measurements in the two groups after applying the control or treatment condition, respectively, permits the **causal conclusion** that the effect is due to the treatment. It is not necessary to perform pretests, or, in view of a possible **reactivity** of such pretests, they even should not be used at all. This reactivity might show up in a **sensitization**, i.e. subjects respond to succeeding measurements in a more sensitive way, or in a **resistibility**, i.e. subjects respond to succeeding measurements in a less sensitive way.

Controlled study: A study, where a **control of extraneous variables** is exerted.

Controlled trial: Synonym for **controlled study**.

Controlled variable: Synonym for **independent variable**.

Control of extraneous variables: By using certain techniques of **global control** (e.g., by **randomization**, **constancy** or **covering**) or **local control** (e.g., by **matching** or **blocking**) it is possible to control extraneous variables, i.e. to neutralize their influence.

Control of substrata: See **stratification**.

Control of the dependent variable: By means of an **experimental design** one wants to find out whether a **dependent variable** shows a systematic change if an **independent variable** is changed systematically.

Control treatment: The experimental condition used in a **control group**.

Control variable: Synonym for **covariate**.

Convenience sample: A **sample** from a **population** which has been formed solely because of its easy availability.

Cooperative study: Synonym for **multicentre study**.

Correlated groups: Groups, where the dependences between the measurements can make the interpretation of outcomes more difficult. An example is the **before-after design**. Another example are groups, between which an interchange of information takes place. Finally, correlated groups are generated if one or more subjects are used in more than one group.

Correlated samples: Synonym for **correlated groups**.

Correlational design: See **correlational study**.

Correlational research design: Synonym for **correlational study**.

Correlational study: Two **dependent variables** are measured at the same sample of subjects and a correlation coefficient is calculated, i.e. a measure of the degree of linear dependence of the two variables. A high coefficient might be solely due to the fact that both variables exhibit a high dependence with a third variable though they are not connected by a causal relation themselves. If the third variable was known and if a sample of subjects was considered, where this third variable is kept constant, no relation between the two first variables would be observed. Therefore, also the term **illusory correlation** is used. Similarly, it might happen that no relation between the two first variables is observed, but that such a relation would be found, if a third unknown variable was kept constant. Thus, the results of correlational studies can give no information about the true relations between variables. In particular, they do not permit **causal conclusions**.

Correlative relation: If a linear relationship is found between two variables, this is called a correlative relation. This is not necessarily a **causal relation**, if one cannot rule out that it is an **illusory correlation**.

Counterbalancing: If each subject is exposed to more than one experimental condition and if the same chronological order of the conditions is used for each subject, one cannot rule out that effects on the **dependent variables** are found which are not caused by the immediately preceding condition but which might be due to **carry-over effects** of former conditions. Similarly, such effects of the chronological order of the conditions can have the consequence that existing effects of single conditions cannot be detected. In order to control effects of chronological order often the technique of counterbalancing is used, where either all possible orders (**complete counterbalancing**) or a selection of possible orders (**incomplete counterbalancing**) of the experimental conditions is assigned to different subjects or groups of subjects, respectively. Because the order of the conditions is determined in a systematic and not in a random way when using counterbalancing, **causal conclusions** can be made implausible by providing **alternative explanations**. A particular problem is that **asymmetric carry-over effects** cannot be ruled out quite often. Problems in interpreting outcomes occur not only in case of incomplete but also in case of complete counterbalancing, e.g., in case of **crossover designs**. The term counterbalancing should not be mixed up with the term **balancing**.

Covariate: This is a **dependent variable** which is recorded in addition and for which one assumes that it has an influence on the dependent variable in which one is interested. This means that not only the considered **independent variables** but also the covariate influence the dependent variable. By isolating the influence of the covariate on the dependent variable by means of statistical methods (analysis of covariance) one tries to achieve that the influence of the independent variable on the dependent variable becomes more distinct. In most cases unrealistic assumptions with respect to the properties of the covariate have to be made, for instance that one has measured it without error, that the relation between covariate and dependent variable is strictly linear and that this relation is the same for all **levels** of the independent variable. In particular, if a pretest is considered as a covariate and the corresponding posttest as a dependent variable, the above assumption of an error-free measured covariate seems implausible.

Covariation: Synonym for **concomitant variation**.

Covering: Technique of **global control** where known **extraneous variables** are covered for all subjects by suitable stimuli. E.g., the noise of cars passing a street near a laboratory, where an experiment takes place, might be covered by a constant noise which is introduced in the laboratory.

Criterion variable: Synonym for **dependent variable**.

Critical case sampling: Critical case sampling is a **purposive sampling**, where only such **sample units** are used for which it is known, in general, from preceding studies, that outcomes for these sample units allow a generalization to the population.

Cross-classification: Cross-classification is present in a **factorial design**, if each **level** of each **factor** is combined with at least two levels of each other factor. A **complete cross-classification** is present, if measurements for all combinations of factor levels are recorded. Otherwise, we have an **incomplete cross-classification**.

Cross-classified design: A **factorial design**, in which all **factors** exhibit a **cross-classification**.

Cross-cultural study: A study which is performed simultaneously in different cultures, i.e. as a rule in different countries, according to a common **protocol**.

Crossed-factor design: Synonym for **factorial design**.

Crossed treatments: Two or more treatments which are applied to the same subjects either simultaneously (**factorial design**) or sequentially (**crossover design**).

Cross-level inference: Conclusions from results for data which were obtained at one level to results for data which were obtained on another level. An example would be an inadmissible conclusion from results for **aggregates** to results for sample units. This has as a consequence the so-called **ecological fallacy**.

Crossover design: Experimental designs which result from using the control technique of **counterbalancing** are also called crossover designs. In the most simple version of such a design we have two experimental conditions (A and B) and to the two possible chronological orders (**sequences**) A_1B_2 and B_1A_2 two independent samples of subjects are assigned. After each of the four conditions A_1, B_2, B_1, and A_2 has been applied, the **dependent variable** is measured. Because of a possible **asymmetric carry-over effect** it is not advisable to consider the measurements after A_1 and A_2 or after B_1 and B_2, respectively, as equivalent. **Causal conclusions** are only possible by comparing the measurements after A_1 and B_1. By this the application of the conditions A_2 and B_2 and the performance of the corresponding posttests becomes unnecessary. However, if one is interested just in studying the possible existence of **carry-over effects**, the design above should be extended to **Balaam's design** by enclosing two additional samples with the sequences A_1A_2 and B_1B_2. Crossover designs are also termed designs of type **COD (*t*, *p*, *s*)**. Here, COD stands for crossover design or **change-over design**, *t* for the number of experimental conditions, *p* for the number of **periods** and *s* for the number of sequences and at the same time for the number of samples. The above crossover design is of the type COD (2, 2, 2), while its extension to Balaam's design is of the type COD (2, 2, 4).

Crossover rate: Portion of those patients in a **clinical trial** which are subsequently assigned to a treatment condition which is different from the condition required by the random allocation.

Cross partition: Formation of **strata** according to the levels of two or more characteristics.

Cross sectional study: In contrast to a **longitudinal study**, only one measurement of each **dependent variable** is recorded at one point of time or within a fixed small interval of time for all subjects of a sample. One tries to draw **causal conclusions** from the observed correlations. Here, the problem of **illusory correlations** arises.

Crucial experiment: An experiment, the outcomes of which should permit a definitive decision between two or more incompatible hypotheses.

Cumulative effect: If a sequence of two or more equal or different experimental conditions is applied in a **within-subjects design** to a subject, it is, starting with the second condition, no longer possible to decide whether an observed effect is solely due to the condition under consideration or also to **carry-over effects** of preceding conditions.

Cyclic trend: See **trend**.

Dark room effect: If subjects believe that they are not being observed, i.e. if no social sanctions are expected, they often behave differently than in situations with social control. If a situation guaranteeing anonymity is provided for, one can try to avoid effects of **social desirability responding** or, at least, to diminish these effects.

Debriefing: Explanation of the true object of an experiment to the participants after the experiment has been performed.

Dehoaxing: A kind of **debriefing**, if the participants of an experiment were wrongly informed about the object of the experiment before the study.

Delayed causation: If a treatment is applied to a subject, it is possible that the effect of the treatment is not observed at once (**abrupt change**) but instead increases gradually to a maximum (**gradual change**). It is also possible that the effect is observed only after a certain known or unknown time interval (delayed causation).

Deliberate sampling: Synonym for **purposive sampling**.

Demand characteristics: Actual or suspected hints which a subject gets about the nature and object of a study and which might influence the behavior of the subject.

Dependent samples: Synonym for **correlated groups**.

Dependent variable: Synonym for **effect variable**, i.e. a variable, for which one supposes that it is influenced by an **independent** or **causal variable**.

Descriptive research: Compare **naturalistic observation study**.

Desensitizing: A kind of **debriefing** used in order to help subjects to appropriately cope with the experiences they themselves made in an experiment.

Design of experiment: Synonym for **experimental design**.

Detection bias: Synonym for **ascertainment bias**.

Developmental design: Case study, where the development of a subject is observed.

Deviant case analysis design: A **quasiexperimental design**, where one tries to identify subjects in a **retrospective study** which deviate from the majority of the other subjects. The researcher compares the values for one or more **dependent variables** for the deviant subjects and tries to formulate hypotheses about possible causes of the deviations. Due to the missing **randomization** it is not possible to draw **causal conclusions**.

Diachronic study: Study, where events are recorded which occur in the course of time.

Diagonal square: Particular **Latin square**, where the treatments in the main diagonal and the other (parallel) diagonals are equal as it is demonstrated in the following example.

C	B	A	D
D	C	B	A
A	D	C	B
B	A	D	C

Difference score: The difference between the values of a **posttest** and a **pretest**. This is a particular case of a **gain score**.

Differential attrition: This occurs, if **attrition** is different for the different treatment conditions.

Differential carry-over effect: Synonym for **asymmetric carry-over effect**.

Differential effect: Synonym for **simple effect**.

Differential mortality: Synonym for **differential attrition**.

Differential transfer: Synonym for **asymmetric carry-over effect**.

Diffusion effect: Synonym for **diffusion of treatments**.

Diffusion of treatments: Exchange of information between subjects who have been assigned to different experimental conditions. Compare also **imitation of treatment**, **compensatory rivalry**, **John Henry effect**, and **resentful demoralization**.

Direct assay: See **bioassay**.

Direct relationship: Simultaneous increase or decrease of the values of two variables.

Direct sampling: Selection of a **sample** of subjects from a **population** without knowing any characteristics of the subjects.

Direct selection: See **selection**.

Direct treatment effect: Treatment effect which is caused by the considered treatment alone and is not modified by **carry-over effects**, **confounding** or **reactivity** of preceding treatments.

Disturbance factor: Synonym for **extraneous variable.**

Domain sampling: Selection of items, e.g. questions for a questionnaire, from a certain field of knowledge, a so-called domain.

Dorfman scheme: A proceeding which is described under **limited data collection**, where at a first level several subjects are tested together, and only, if a suspicious outcome is observed, the corresponding subjects are identified at a second level by separate tests (Dorfman, 1943).

Dose modification: In a **phase III study** of a longer duration it might be necessary for medical reasons to alter the dose of a drug which was assigned to a patient for a shorter or longer time. In the **protocol** it should be fixed in advance when and how such dose modifications are to be performed.

Dose-ranging trial: A **clinical trial**, to find the appropriate size of the dose of a drug for the initial and the subsequent applications.

Dose-response curve: See **bioassay** and **dose-response experiment**.

Dose-response experiment: In such an experiment, the **levels** of the **independent variable** correspond to the doses of a drug. To each of these levels a sample of subjects is randomly assigned (**randomization**) and values of a **dependent variable** are recorded. As a result a **dose-response curve** is obtained which describes the functional relationship between dose and effect.

Double balanced incomplete block design: A **balanced incomplete block design**, where not only the pairs of treatments but also the triplets of treatments occur the same number of times. In the following example, each of the $b = 4$ rows corresponds to a **block** with $k = 3$ subjects, and A, B, C, and D correspond to the $t = 4$ treatments of which each occurs $r = 3$ times. While each pair of treatments occurs in $c = 2$ blocks, each triplet of treatments occurs in only one block.

ABC	ABD	ACD	BCD

Double blindfold experiment: See **blinding**.

Double-blind study: See **blinding**.

Double block design: If **blocking** is being used and the **blocks** are formed with respect to two **extraneous variables** instead of only one, a double block design results. **Latin squares** are one example.

Double block design with nested block factors: A **double block design**, where the **levels** of one of the **block factors** are partitioned into subgroups (of one or more levels), where each subgroup can be combined only with a certain level of the other block, resulting in a **hierarchic block structure**. Also see **nested design**.

Double confounding: This is present if a **confounding** with two **extraneous variables** is present for a **factorial effect** in a **factorial experiment**.

Double dummy technique: If the treatments in a **clinical trial** differ very much, it is difficult to perform **double-blind studies**. Sometimes it might be possible to use double dummy techniques. If, e.g., a standard drug and a new drug are to be compared and if only an oral application is possible for the standard drug while at the same time only an intravenous application can be used for the new drug, a group of patients would get the standard drug and in addition intravenously a **placebo** while another group of patients would get the new drug and in addition orally a placebo.

Double grouping: The presentation of the outcomes of a **factorial experiment** with two **factors** in a rectangular scheme, where the rows (columns) correspond to the **levels** of the first (second) factor. This corresponds to a **two-fold classification** of the outcomes.

Double inspection: The same characteristic is measured at the same **sample** of subjects at two different points of time.

Double-masked study: Synonym for **double-blind study**.

Double observation: The same characteristic is recorded for one subject at two different points of time.

Double sampling: See **two-stage sampling**.

Doubly censored data: If neither the time of the outbreak of a disease nor the time of the death of a patient for which the survival time is to be determined is known, we have doubly censored data.

Drift: Denotes systematic changes, e.g., **trends**, which are observed in **within-subjects designs** with constant outer conditions.

Dropout: A subject for which, due to **experimental mortality,** not all measurements are available.

Dropout rate: Synonym for **attrition rate**.

Dual block design: A dual block design with respect to a **block design** is obtained by interchanging the role of treatments and **blocks**. E.g., consider the block design

AB	AC	AD	BC	BD	CD

with 6 blocks, 4 treatments, and 2 subjects in each block. The corresponding dual block design

ABC	AEF	BDF	CEF

consists of 4 blocks, 6 treatments, and 3 subjects for each block. Formally, the dual block design is constructed by interchanging rows and columns in the **incidence matrix**.

Duplicated sample: The same characteristic is measured at the same sample of subjects by two different persons.

Dummy combination: If we split up the set of **levels** of a **factor** artificially in order to obtain a **cross-classification** with two factors, combinations of levels can result which cannot be realized. E.g., consider a study investigating whether it is better to remove gallstones by means of an operation, to dissolve them chemically or simply to put the patient on an appropriate diet. This yields a factor with three levels. However, one could also consider the factor "surgical removal" with the levels "diet" and "operation" and the factor "chemical removal" with the levels "diet" and "dissolution". In this case the combination of the levels "operation" and "dissolution" is a dummy combination because it cannot be realized.

Dummy experiment: A **sample** of subjects is randomly split up into two subsamples of different size and is exposed to only one experimental condition. From the outcomes we might obtain important information with respect to the actual experiment concerning the necessary **sample size**, the nature and size of **blocks**, etc.

Dummy treatment: Synonym for **placebo**.

Dummy trial: Synonym for **dummy experiment**.

DV: Abbreviation for **dependent variable**.

Dynamic population: In many cases, where a **sample** is selected from a **population**, the population cannot be considered as a **static population** which does not change in time. Due to aging and other changes of the subjects, the entrance and leaving of subjects, a dynamic population must be assumed which is altered even during the performance of the study. Therefore, any statements about a population concern always a population that does no longer exist in its original composition or which has never existed, respectively.

Ecological correlation: A correlation, i.e. a measure of a linear relationship, between two **dependent variables**, which is based on grouped data, i.e. on averaged values from **aggregates** and not on outcomes from **sample units**. In general, it is not allowed to conclude from the existence of such an ecological correlation the existence of a correlation for the sample units (**ecological fallacy**). See also **cross-level inference**.

Ecological fallacy: See **cross-level inference** and **ecological correlation**.

Edgington design: Single-subject design where to a single subject an **experimental design** is randomly allocated (**randomization**) according to an idea of R. A. Fisher (Edgington, 1995, Chapter 12). **Causal conclusions** are possible, if a **double-blind study** is used.

Effect: Difference in the values of a **dependent variable** caused by the difference of the levels of one or more **independent variables**.

Effect modifier: An **extraneous variable**, which exhibits an **interaction** with the studied **independent variable**.

Effective sample size: The sample size of that **sample** which results after all the subjects have been excluded from the study (according to strict criteria which were fixed in advance) whose outcomes would bias the evaluation. The excluded subjects might be **dropouts**, subjects with **missing observations**, subjects for which one finds subsequently out that they do not meet the **inclusion** and **exclusion criteria** or subjects which have exhibited an insufficient **compliance.**

Effect of selection: Synonym for **selection effect**.

Effect variable: Synonym for **dependent variable**.

Efficacy: The **effect** of a treatment in comparison with a control condition in the ideal situation that after a **random allocation** of the subjects no **dropouts** or **missing observations** have occurred.

Efficacy population: Synonym for **per protocol population**.

Elaboration: Attempt to find out whether a correlation between two variables is an **illusory correlation** by keeping a third variable constant.

Elementary balanced incomplete block design: A **balanced incomplete block design** which is no **composite balanced incomplete block design**.

Elementary design: Synonym for **completely randomized design**.

Eligibility: A subject is eligible for a **clinical trial** if all **inclusion** and **exclusion criteria** are met.

Elimination: This is a **global control** technique where an **extraneous variable** is eliminated from the experimental situation.

Endogenous variable: A **factor** which influences a system in a controlled way, e.g., because it is an **independent variable** or a **constant factor**. However, also the **dependent variables** are considered as endogenous variables. See also **exogenous variable**.

Endpoint: Unequivocally defined outcome after which a subject leaves a study. In **clinical trials** this is typically either death or complete recovery of a patient. Sometimes endpoint is used as a synonym for **dependent variable**. See also **multiple endpoint**, **primary endpoint**, **subjective endpoint**, and **surrogate endpoint**.

EPBCD: Abbreviation for **extra period balanced crossover design**.

Epidemiological study: Serves for studying the relation between a disease and those **factors** which possibly have an influence on the disease.

Equal probability sampling: Synonym for **simple random sampling**.

Equipotent dose: By use of a standard drug a certain effect is found for a certain dose. That dose of a new drug by which one gets the same effect is called equipotent dose.

Equireplicate block design: A **block design**, in which each of the considered treatments occurs the same number of times.

Equivalent dose: Synonym for **equipotent dose**.

Equivalent time samples design: A sequence of time intervals is fixed for only one group of subjects. Different treatment conditions are randomly assigned to these time intervals (Cook and Campbell, 1979, pp. 377-378). Consider as an example a store where 25 of 50 days are randomly selected, during which a certain adjustment of the air-conditioning plant is being used while for the other 25 days another adjustment has been chosen. The daily turnover is chosen as a **dependent variable**.

Error of central tendency: See **response bias**.

Error of leniency: See **response bias**.

Error variance: That portion of variation in the values of a **dependent variable**, which cannot be explained by the effects of known **independent** or **extraneous variables** but which is due to unknown or not recorded extraneous variables.

Evaluable patient population: Synonym for **per protocol population**.

Evaluable patients: All patients in a **clinical trial** which are not excluded from the final evaluation. The number of these patients is given by the **effective sample size**.

Evaluation apprehension: Synonym for **social desirability responding**.

Event history data: Subjects go through a series of states, e.g., states of a disease. The chronological order of the states, the sojourn times in the states and maybe the transition times from one state to the next one can be recorded.

Exact experimental design: An experimental design which is completely described and realizable, i.e. which does not, e.g., contain any **dummy combinations**.

Examiner bias: Synonym for **experimenter effect**.

Exclusion criteria: Exclusion criteria describe properties of those subjects which should not be included in a study.

Exogenous variable: A **factor** which influences a system from the outside in an uncontrolled way. In other words, it is an **extraneous variable** which is not controlled.

Expectancy control group: A **control group** which is used in order to control the **Rosenthal effect**. A certain expectancy with respect to the outcome of the experiment for the subjects in this group is transmitted to the experimenter.

Expectancy effect: Synonym for **Rosenthal effect**.

Experiment: An empirical method which permits **causal conclusions**.

Experimental conditions: All **factors** which can influence **dependent variables**. These are not only the **independent variables** under consideration but also **extraneous variables** or **constant factors**.

Experimental contamination: Synonym for **reactivity**.

Experimental control: The experimenter can fix arbitrarily when, where, and how events take place, which themselves are defined in an arbitrary way. Further, the environmental conditions are arbitrarily fixed and the experimenter can repeat the experiment at arbitrarily chosen points of time in just the same way as before. Finally, the experimenter can vary in a systematic way the experimental conditions, to be able to detect changes in the **dependent variable** which are caused by these conditions. Without experimental control no **causal conclusions** can be drawn because of possible **alternative explanations**. According to McGuigan (1978, p. 147) experimental control means **independent variable control** as well as **extraneous variable control**.

Experimental demand: Tendency of subjects to behave according to observed or suspected hypotheses or expectancies of experimenters.

Experimental design: This is mainly defined by stating the **independent variables** and their **levels**, the permitted combinations of the levels, the chronological order of the influence of the independent variables, the **dependent variables** and the points of time when the respective dependent variables are recorded. When the independent and dependent variables are stated, it is necessary to state in addition their **operational definitions**. One has to mention which combinations of the levels are to be used for different samples and which at the same sample. Further, the size of each sample should be given. Finally the control techniques used, e.g., **randomization** or **blocking**, should be described. Only outcomes from carefully planned experimental designs allow **causal conclusions**.

Experimental error: This is caused by differences between the subjects, **extraneous variables** which have not been eliminated and **measurement errors**. It causes that variation of the values of the **dependent variables** which remains, after the influence of all known **factors** has been taken into account.

Experimental group: Synonym for **treatment group**.

Experimental mortality: Experimental mortality of a subject means that the data of this subject participating in an experiment are, completely or partly, no longer available from a certain point of time on. This loss of data does by no means has to have been caused by the actual death of the subject. It might just as well be due to a failure of a measuring device, to a mistake of the experimenter, to insufficient **compliance** of the subject, to illness of the subject etc.

Experimental study: Synonym for **experiment**.

Experimental unit: In most cases the subject itself is the experimental unit. However, if subjects are not to be studied isolated from each other, e.g., when considering therapy groups, families, school classes, inhabitants of a house etc., more complex experimental units arise. Thus any **causal conclusions** can be drawn and formulated only with respect to these larger experimental units.

Experimenter bias: Synonym for **experimenter effect**.

Experimenter effect: All effects which the experimenter has on the **dependent variable** due to his or her characteristics (e.g., age, gender, body length, figure, voice, clothes etc.) or to his or her behavior. In particular, expectancies of the experimenter might be reflected in his or her behavior and thus result in that experimenter effect which is known as **Rosenthal effect**.

Experimenter expectancy: See **Rosenthal effect**.

Experimentum crucis: Synonym for **crucial experiment**.

Explained variance: That portion of the variation of the values of the **dependent variables**, which can be explained by the variation of the **independent variables**.

Explanatory approach: In contrast to the **intention-to-treat analysis** or the **treatment received analysis** only those patients are considered in a **clinical trial**, who received the treatment condition which was assigned to them in accordance with the **protocol**. When the outcomes are being interpreted, **selection effects** cannot be ruled out.

Explanatory research: Each kind of research with the object to detect **causal relations** between variables and to prove their existence.

Explanatory trial: Clinical trial in which it is not the object to find out whether a treatment has an effect but rather to determine the mechanism of this effect.

Explanatory variable: Each variable which can have an influence on a **dependent variable**.

Exploratory experiment: According to Robinson (1976, p. 173) this is an experiment in which a **control group** is compared with a **treatment group** to answer the following question: Does there exist any effect of the **independent variable** irrespective of the question which effects are observed for different **levels** of the independent variable? In McGuigan (1978, p. 75) the term exploratory experiment is used in a more conventional sense. According to this definition it is an experiment which is performed in a field where the base of knowledge is small and where one tries to look for possible effects having only speculations and more or less empirical evidence. The opposite to an exploratory experiment in this sense would be a **confirmatory experiment**.

Exploratory study: A study which, though it should have a clear and precise object, is not performed for testing certain hypotheses but is to be used to generate new hypotheses.

Ex post facto design: Synonym for **quasi-experimental design**.

Exposure factor: Synonym for **risk factor**.

Extensive sampling: In general, the **sample size** is very small in comparison with the population size. If, however, a considerable portion of the **population** is contained in the **sample** this is called extensive sampling.

External validity: The higher the external validity of a study, the higher the range, where the detected effects are valid, i.e. the higher the generalizability of the found results to larger populations of subjects or to more general situations.

Extraneous variable: Though this is a variable which has an influence on the **dependent variable** under consideration, we are not interested in any effects of this **causal variable**. The influence which is exerted by an extraneous variable on the dependent variable might either have a direct effect or might be due to an interaction between the extraneous and the **independent variable** under consideration. In the latter case the difference of the effects of two levels of the independent variable would be different for different levels of the extraneous variable.

Extraneous variable as independent variable: As a possible **control of extraneous variables** it is sometimes proposed to use known **extraneous variables**, e.g., age, gender or body weight, as **independent variables** in the **experimental design**. Because these will be usually variables for which no random assignment (**randomization**) of the levels to the subjects is possible, **causal conclusions** with respect to these variables are not possible, because **selection effects** cannot be ruled out. Obviously, the proceeding is similar to **matching** or **blocking**.

Extraneous variable control: Control of the **extraneous variables** so that they cannot exert a systematic influence on the **dependent variables** which could be erroneously ascribed to the **independent variables**. I.e., extraneous variable control should prevent a **confounding** of independent and extraneous variables. It is one component of **experimental control**.

Extraneous variance: That variation in the values of a **dependent variable** which is caused by **extraneous variables**.

Extra period balanced crossover design: A **balanced crossover design**, where the last **period** is repeated. An example with four periods and six sequences is given in the following. Here, each chronological order of two equal or different treatments of altogether three treatments occurs exactly two times.

A	B	C	A	B	C
B	C	A	C	A	B
C	A	B	B	C	A
C	A	B	B	C	A

Extreme groups: See **statistical regression**.

Extreme values of the independent variable: When selecting the **levels** of an **independent variable**, sometimes two levels are chosen which are as far apart from each other as possible. By this it is hoped to obtain a large difference in the **dependent variable**, thereby, facilitating a **causal conclusion**. However, this object might not be achieved if the relation between the independent and the dependent variable is not monotonic. This is the case, e.g., if for the extreme values nearly equal values of the dependent variable result while for intermediate values much larger values are obtained. If the choice of the extreme values is performed by means of a **manifest variable**, measurement errors of this variable can cause a wrong selection of extreme values.

Factor: Synonym for **independent variable**.

Factorial design: Experimental design with two or more **independent variables**, where it might be impossible to obtain measurements for all combinations of **levels**.

Factorial effect: An effect on a **dependent variable** which is caused by a **factor** in a **factorial experiment**.

Factorial experiment: Synonym for **complex experiment**.

Factor level: Synonym for **level**.

Fallback quasiexperiment: According to Cook and Campbell (1979, p. 134) this is a **quasiexperimental design** which is scheduled in addition when planning an experiment in order to improve the interpretability of the outcomes of the experiment. It might be possible, e.g., that the outcomes of an experiment do not permit **causal conclusions** due to **differential attrition**. If this possibility cannot be ruled out before the experiment is performed, one could try to make implausible at least some possible **alternative explanations** by recording additional measurements and by considering additional groups of subjects.

Fan-spread model: In **observational studies** often a **fan-spread pattern** is observed, where small differences in the measurements of subjects become considerably larger in the course of time even without any intervention from the outside. A possible explanation for this is the fan-spread model which states that also in seemingly homogeneous groups the healthier, stronger, more intelligent etc. subjects have the largest potential for a positive change.

Fan-spread pattern: See **fan-spread model**.

Feasibility study: Synonym for **pilot study**.

Fibonacci dose escalation scheme: Sometimes recommended procedure for establishing the **maximum tolerated dose** of a pharmacon during a **phase I study**. The procedure consists of the application of an increasing sequence of dose levels to subjects where the rate of increase decreases from dose to dose. The initial dose level d_1 is derived by a conservative consideration of the outcomes of animal experiments or from previous human experience. If for a certain dose level an occurrence of toxicity is observed that is unacceptable according to a given criterion this event is called in the following a dose-limiting toxicity (DLT).

The Fibonacci numbers are defined as a sequence of integers, the two first terms of which are set to 1, while each following term is the sum of the two preceding ones, i.e., 1, 1, 2 = 1 + 1, 3 = 1 + 2, 5 = 2 + 3, 8 = 3 + 5, etc. For the present problem the sequence of the ratios of two succeeding Fibonacci numbers is considered, i.e., 1 = 1/1, 2 = 2/1, 1.5 = 3/2, 1.667 = 5/3, 1.6 = 8/5, 1.625 = 13/8, etc. This sequence converges to $2/(\sqrt{5}-1)$ = 1.618 (golden ratio). Because, obviously, the rates of increase for this "true" Fibonacci sequence form no strictly decreasing sequence, "modified" Fibonacci sequences are considered, defined, e.g. by 1, 2, 1.67, 1.5, 1.4, 1.3, 1.3, 1.3, ... This means that a sequence of dose levels where the rate of increase decreases from dose to dose is defined by d_1, $d_2 = 2d_1$, $d_3 = 1.67d_2$, $d_4 = 1.5d_3$, $d_5 = 1.4d_4$, $d_6 = 1.3d_5$ etc. (Note that there is always an increase of the factor from d_1 to d_2!).

A corresponding escalation scheme might have the following form: First, three patients are exposed to the initial dose level d_1. If for no patient a DLT occurs, three new patients get dose level d_2. If for two or three patients a DLT occurs, the dose escalation scheme discontinues. If a DLT occurs for only one patient, three additional patients get dose level d_1. If DLT occurs for none of the three additional patients, three new patients get dose level d_2, otherwise the dose escalation scheme discontinues. If one of the two cases occurs, where dose level d_2 is given to three new patients the same procedure as for dose level d_1 is performed, etc. If one stops the trial at a certain dose level, the maximum tolerated dose is the dose of the preceding dose level. If the trial already stops for dose level d_1, a completely new trial with a more conservative initial dose level has to be performed to estimate the maximum tolerated dose.

Field experiment: A **field study**, where, in contrast to common practice, a **control of extraneous variables** takes place, in particular by **randomization**. **Causal conclusions** are possible.

Field study: Field studies are scientific investigations which take place outside laboratories under realistic conditions. While it is an advantage of such studies that they are nearer to reality, a decisive disadvantage is that a **control of extraneous variables**, in particular by **randomization**, is only possible in rare situations. Therefore, **causal conclusions**, in general, are not admitted.

Field trial: Synonym for **field study**.

First-order correlation: Correlation, i.e. a linear relation, between two variables, while a third variable is kept constant. This might be considered, e.g., in order to avoid an **illusory correlation.**

First-stage unit: See **two-stage sampling**.

Fisher block design: A **randomized block design**, where each treatment occurs exactly once in each **block**.

Five-point assay: Specific design in drug studies to compare a new drug with a standard drug. A first group receives a **placebo** or no treatment, a second group receives a dose of the standard drug, a third group receives a twice as high dose of the standard drug, a fourth group receives a dose of the new drug, and a fifth group receives a twice as high dose of the new drug.

Fixed-effects model: Synonym for **fixed model**.

Fixed factor: See **fixed model**.

Fixed model: In this model one assumes that the **levels** of the **independent variables** have been fixed in the way described under **purposive manipulation of the levels of the independent variable**.

Fixed sample: A fixed sample is given in the case of **repeated sampling**, if the **sample** contains at each point of time the same subjects.

Fixed sample size: See **sequential sampling**.

Floor effect: Measuring devices, in particular those for measuring behavior, produce only values for measuring a certain range, i.e. there exist a smallest and a largest possible measurement value. Of course, this does not mean, that it is impossible to develop devices by which also behavior can be measured which is outside the range of a given device. If for several subjects measurements are recorded which are equal to the smallest possible value, these subjects cannot be distinguished with respect to the recorded **dependent variable** and a so-called floor or **basement effect** results. This effect causes an increased inaccuracy of measurement which is the larger the nearer the measurement is to the minimum possible value. An analogous effect is observed if a measurement is nearer to or even equal to the largest possible value. In this case, a **ceiling effect** results.

Follow-up study: After a study has been performed, subjects are studied over a long period in appropriately chosen time intervals with respect to the occurrence of certain well-defined events.

Forward-looking study: Synonym for **prospective study**.

Four-point assay: Specific four-group design in drug studies for comparing a standard drug with a new drug. For this, two groups receive two different doses (D_1 and D_2) of the standard drug and two other groups the same doses (D_1 and D_2) of the new drug.

Fractional replication: Incomplete factorial design, where a fixed proportion of the possible combinations of the factor levels is considered, e.g., half of the possible combinations in a **half-replicate design**.

Frailty: Term used for causes of differences between the subjects in a **population** which are either not known or are difficult to measure. Therefore, frailty is responsible for a part of the **error variance**. Frailty might also be the cause of an **illusory correlation** which disappears if frailty is kept constant. E.g., consider a population consisting of families. As a dependent variable we consider the lifetime of a subject. It seems reasonable to assume that the lifetimes of the members of a specific family are influenced by common unobserved risk factors. This influence is considered as a frailty, i.e. there exists a value of this frailty specific to each family.

Frame: Complete list of subjects in a **population**, from which a **representative sample** is to be selected.

Friedman's urn model: If subjects enter a study sequentially and have to be assigned to a **treatment** or **control group** at once, different ways how to guarantee that the sizes of the two groups do not differ too much are conceivable, even though a **systematic design** is avoided. One way is the **permuted block design**. Another way is Friedman's urn model following an idea by Friedman (1949). For this $a \geq 1$ cards with the inscription "treatment" and the same number of cards with the inscription "control" are prepared. The cards are shuffled and laid into a box. If the first subject enters the study a card is drawn randomly (**sampling without replacement**) and the subject is assigned to the corresponding group. After that $(1 + \alpha)$ additional cards with the same inscription as the drawn card and β additional cards with the other inscription are laid into the box and the cards are shuffled again. Here, we assume $\alpha \geq 0$ and $\beta \geq 0$. For the second subject again a card is drawn (sampling without replacement) and again $(1 + \alpha)$ and β additional cards are laid into the box, etc. If **random allocation** should be more important than equal sample sizes, a high value of a is chosen, otherwise a small value of a is fixed. The larger β in comparison to α, the more similar sample sizes result. A generalization to more than two experimental conditions is achieved by assuming $a \geq 1$ initial cards for each condition. Then, for the selected condition $(1 + \alpha)$ and for each of the remaining conditions β cards are added.

Gain score: This is a score which in most cases is calculated from **pretest** and **posttest** values. It should reflect the true **effect** of a treatment. An example for a gain score is the **difference score**.

Gambler's fallacy: If data are ordered in time or space, observers tend to detect laws which do not exist in reality. In particular, if certain diseases or accidents accumulate in a short time interval or in a small geographic region, it is near at hand to assume a common cause, even if only a random accumulation has occurred.

Generalizability: Synonym for **external validity**.

Generation effect: Synonym for **cohort effect**.

Global control: Control of extraneous variables simultaneously for all subjects which participate in an experiment. Examples are **randomization** or **covering**.

Gold standard trial: Clinical trial, by which a standard treatment (so-called gold standard), a new treatment and possibly a sham treatment are compared.

Gradual change: See **delayed causation**.

Greco-Latin cube: Generalization of a **Greco-Latin square** for the control of four **extraneous variables**. As an example with three treatments A, B, and C the three layers of a Greco-Latin cube are given below. It is obvious that not only in each row and column of each of the three depicted horizontal layers each Latin and Greek letter occurs exactly once and similarly each combination of a Latin with a Greek letter, but that these properties also hold for all vertical layers.

Aα	Bγ	Cβ
Bβ	Cα	Aγ
Cγ	Aβ	Bα

Cβ	Aα	Bγ
Aγ	Bβ	Cα
Bα	Cγ	Aβ

Bγ	Cβ	Aα
Cα	Aγ	Bβ
Aβ	Bα	Cγ

Greco-Latin square: Generalization of a **Latin square** with three **extraneous variables** which have to be controlled. An example of a

Greco-Latin square with three treatments and three extraneous variables, each with three **levels**, is given below. Here, the Latin letters A, B, and C correspond to the three treatments, the levels of the first extraneous variable to the rows, the levels of the second extraneous variable to the columns, and the levels of the third extraneous variable to the Greek letters α, β, and γ. Each Latin and each Greek letter occurs exactly once in each row and column. Each combination of a Latin and a Greek letter occurs only once.

Aα	Bγ	Cβ
Bβ	Cα	Aγ
Cγ	Aβ	Bα

Greco-Latin-square counterbalancing: For this kind of **incomplete counterbalancing** the respective number of **levels** for two **independent variables**, the number of points of time, and the number of considered chronological orders, i.e. also the number of experimental groups are identical. Each level of each of the two independent variables occurs exactly at one position and each level of one independent variable is combined exactly once with each level of the other independent variable. In the following table A, B, C, and D correspond to the levels of one independent variable and α, β, γ, and δ to the levels of the other independent variable. The rows correspond to the experimental groups, the columns to the points of time. Because **asymmetric carry-over effects** are possible, in general, **causal conclusions** are not permitted.

Aα	Bδ	Dγ	Cβ
Cδ	Dα	Bβ	Aγ
Bγ	Aβ	Cα	Dδ
Dβ	Cγ	Aδ	Bα

Grid sampling: Synonym for **lattice sampling**.

Group effect: Those influences on a **dependent variable** which are due to the fact that a subject belongs to a certain subpopulation. If, e.g., a subject belongs to the subpopulation of women, this yields a group effect on dependent variables as height, weight, and income.

Group experiment: A group of subjects which are homogeneous with respect to one or more **block variables** is split up into subgroups which are assigned to different treatments. This corresponds to the **global control** technique described under **constancy**.

Group selection: This refers to the selection of groups of subjects in contrast to the **selection** of single subjects.

Group sequential design: A **sequential design**, where several treatments, corresponding to different **arms of a study**, are compared with each other. Whenever it occurs that for each arm a fixed number of patients has been treated in a **clinical trial**, it is decided on the basis of the outcomes known at that point of time whether the investigation is stopped for single arms or for the whole study.

Habituation: If a subject is exposed to an experimental situation, a habituation might take place with the consequence that its behavior is no longer influenced by the total complexity of the situation but that only responses to specific stimuli are observed.

Half-replicate design: See **fractional replication**.

Halo bias: Synonym for **halo effect**.

Halo effect: Observers tend to give subjects a positive rating for a performance if the same subjects have shown good results with respect to other tasks or if the behavior or even only the appearance of the subjects impresses the observers in a positive way.

Hawthorne effect: By this is meant the effect that the behavior of subjects is altered by their knowledge that they participate in a study.

Healthy worker effect: A particular kind of the **selection effect**. If one assumes that an environmental factor increases the risk of a disease, it is possible that only a relatively small portion of sick subjects is found in that subpopulation of subjects which is exposed to this risk. The reason for this is that sick subjects will no longer expose themselves to the expected risk and will be replaced by healthy subjects sooner or later. In industrial **cohort studies** the healthy worker effect can have two results: First, sick subjects are not engaged. Second, subjects who become sick have a greater risk to loose their job and to be replaced by healthy subjects.

Hello-goodbye effect: This is an effect which can be observed for subjects that have participated in a therapy. Quite often such patients try to signal a high positive therapy effect, consciously or unconsciously. This seeming effect is even increased in its magnitude if these patients try to emphasize their need for help before therapy. As a result of a hello-goodbye effect small positive therapy effects might appear as very important and not existing or even negative effects appear as positive.

Heterogeneity: This is present in a **population** if the subjects differ with respect to a characteristic. If there are no such differences, we have **homogeneity**.

Hidden time effects: These are differences between subsequent measurements in a **within-subjects design** which are not caused by an **intervention** of the experimenter but are due, e.g., to **history**.

Hierarchical design: Synonym for **nested design**.

Hierarchic block structure: See **double block design**.

Historical control: If **control groups** are not produced by a random splitting of an initial sample (**randomization**), but if measurements of control groups from earlier studies are considered instead, the term historical controls is used. Even if these earlier control groups were produced by randomization, **causal conclusions** are not possible because, due to **history**, **selection**, and other effects, the historical control groups do not have to be comparable with the present **treatment groups**.

Historical prospective study: This is either a **retrospective study** or a reanalysis of a **prospective study** which was performed at an earlier point of time. One assumes that all data are available for the prospective study though it is possible that in the original study a different **dependent variable** than in the original study is used as a **primary variable**.

History: A **repeated-measures design** does not guarantee that differences between a **pretest** and a **posttest** corresponding to a treatment are caused by an effect of the treatment. They might rather be due to a change of the environment which was not controlled by the experimenter and which affects the **dependent variable**. Such changes of the environment which are not controlled are called history. A control of history can only be performed by the use of a **control group** without a treatment.

Holdover effect: Synonym for **carry-over effect**.

Homogeneity: See **heterogeneity**.

Hospital controls: See **community controls**.

Household survey: A **survey**, where subjects are interviewed in their household and where it might be that one subject gives information about the other members of the household.

Hyper Greco-Latin square: Generalization of the **Latin square** to the case where 4 **extraneous variables** are to be controlled. For 4 treatments and 4 extraneous variables each with 4 **levels** an example is given in the following. Here, the 4 treatments are assigned to the Latin letters A, B, C, and D, the levels of the first extraneous variable to the rows, the levels of the second extraneous variable to the columns, the levels of the third extraneous variable to the Greek letters α, β, γ, and δ, and the levels of the fourth extraneous variable to the figures 1, 2, 3, and 4. Each Latin letter just as each Greek letter and each figure occurs only once in each row and column. Further, each combination of two letters or of a letter with a figure occurs only once.

Aα1	Bβ2	Cγ3	Dδ4
Bδ3	Aγ4	Dβ1	Cα2
Cβ4	Dα3	Aδ2	Bγ1
Dγ2	Cδ1	Bα4	Aβ3

Hypothesis guessing: If subjects under different experimental conditions form hypotheses about the object of an experiment, this might change their behavior so drastically that existing **causal relations** are not detected or that not really existing effects are found.

Illusory correlation: Such a correlation is present if between two **dependent variables** a relation is observed which is no longer present if a **third variable** is kept constant at the same time. That such illusory correlations might occur has as a consequence that it is not permitted to draw **causal conclusions** on the basis of the outcomes of **correlational studies** because it can never be ruled out that there exist third variables which cause illusory correlations.

Imitation of treatment: A possible consequence of the **diffusion of treatments** might be, e.g., that subjects of a **control group** try to obtain outside the study the same treatment as the subjects of the **treatment group** within the study. In this case it might become impossible to detect a treatment effect.

Impressionistic modal instance model: According to Cook and Campbell (1979, p. 77) this is a proceeding to increase the **external validity**. In order to do so the classes of subjects, situations, and times to which the **causal conclusions** are to be generalized are fixed. For each of these classes at least one specimen is sampled which seems to be typical for this class. Then, the study is performed for these samples.

Impression management: Attempt of subjects to manipulate the impression they make on the experimenter.

Incidence matrix: The incidence matrix indicates for each **block** and each treatment of a **block design** how often the treatment occurs in the block. Consider, e.g., for 4 blocks and 3 treatments the following block design:

AAB	CB	ABC	B

Then, the corresponding incidence matrix is given by the following scheme, where the rows correspond to the treatments and the columns to the blocks.

2	0	1	0
1	1	1	1
0	1	1	0

Inclusion criteria: Inclusion criteria describe the properties which subjects should have to be included into a study.

Incomplete balancing: This is present if one does not assign the same number of subjects to each combination of the **levels** of one or more known **extraneous variables**.

Incomplete block design: A **block design** with at least one **block**, where not all considered treatments occur.

Incomplete counterbalancing: Because the number of required subjects increases very rapidly with the number of experimental conditions in **complete counterbalancing**, in practice only incomplete counterbalancing is often applied. Here, not all the possible arrangements of the experimental conditions are considered, but only a subset of these. In many cases each experimental condition has to occur the same number of times at each timely position in this subset of arrangements. If, here, the number of arrangements is equal to the number of experimental conditions, i.e. if each condition occurs at each timely position exactly once, we have a **Latin square**. In case of a **Greco-Latin square** two **independent variables** with the same number of **levels** are considered, where the number of points of time is equal to the number of levels of one of the independent variables. It is required that each level of each of the two independent variables occurs exactly once at each timely position and that at the same time each combination of a level of one independent variable with a level of the other independent variable occurs exactly once. If it can be realized, it can be additionally required for incomplete counterbalancing that each experimental condition occurs the same number of times before and the same number of times after each other condition (cf. **balanced latin square**). In an experimental design, with the four experimental conditions A, B, C, and D and the sequences ABCD, DCBA, CADB, and BDAC, both requirements would be met because each of the four conditions occurs exactly once at each of the four timely positions and each condition is used exactly once before and once after each other condition. For incomplete counterbalancing it is even more difficult to draw **causal conclusions** than it is for complete counterbalancing.

Incomplete cross-classification: See **cross-classification**.

Incomplete factorial design: Factorial design, where not all possible combinations of the **factor levels** are considered.

Incomplete Latin square: Design which arises by omitting rows and/or columns in a **Latin square**.

Incomplete paired comparison design: Paired comparison design, where at least one pair of subjects is not rated.

Incomplete within-subjects design: According to Underwood and Shaughnessy (1975, p.10, pp.76-83) this is a **within-subjects design**, in which each subject receives each experimental condition only once. Due to this it is no longer possible to control the **progressive error**. **Causal conclusions** are not possible.

Independent censoring: See **non-informative censoring**.

Independent-groups design: See **independent two group design**, though more than two groups might be considered.

Independent trials: Several **trials** with outcomes which do not mutually influence each other. If even only one subject participates at more than one trial, the corresponding trials have to be considered to be dependent, i.e. to be not independent. Dependence between trials might also occur if subjects before participating in a trial are informed about the outcome of another trial.

Independent two group design: A **sample** of subjects is randomly split up into two subsamples (**randomization**) which are assigned to two different experimental conditions. Here, in principle, **causal conclusions** are possible.

Independent variable: Synonym for **causal variable**, i.e. a variable which possibly has an effect on a **dependent variable** or **effect variable**. This effect is to be studied.

Independent variable control: Possibility to vary **independent variables** in a known and intended way. This is one component of **experimental control**.

Indicator variable: Sometimes used as a synonym for **manifest variable**.

Indirect assay: See **bioassay**.

Indirect sampling: Selection of a **sample** of subjects from a **population** after the characteristics of interest were already recorded for the subjects of the population and are known. The sample is then selected from the records of the observations.

Indirect selection: See **selection**.

Individual-difference variable: Synonym for **background variable**.

Individual observation: A single characteristic is recorded only once for one subject.

Individual selection: If the **selection** is related to subjects and not to groups of subjects.

Infertile worker effect: It is well-known that working women have fewer children than housewives. However, from this one cannot conclude that occupation causes infertility, because it is possible that women give up their occupation as soon as they have children or because working women practice a deliberate birth-control in order to be able to work furtheron.

Information bias: Fallacies which are caused by a different **precision** with which the **independent** and the **dependent variables** are registered in different groups.

Informative censoring: This is present if the probability of the occurrence of censored observations in a **censored sample** depends on the respective experimental conditions, e.g., because the treatment of a patient has to be interrupted due to considerable **adverse effects**. In case of such an **interaction** between **independent variable** and censoring mechanism, in general, **causal conclusions** are not possible.

Informative missing value: Missing value, for which the probability of occurrence depends on the magnitude of the actually observed values and/or on the magnitude of the non-observed values.

Informed consent: Explicit agreement of a patient to participate in a **clinical trial** after a detailed information has been given about object and nature of the study, about the method of assignment to the different treatments, about the risks and chances for the patient under consideration, about the kind of measurements used, and about the time schedule of the study including not only the treatments but also the performance of the measurements.

Inhomogeneity: Synonym for **heterogeneity**.

Instrumental variable: Variable which is correlated with an **explanatory variable** but which itself has no direct influence on the considered **dependent variables**.

Instrumentation: This means, e.g., that effects seem to occur only because measurement devices change in the course of time. Such changes might concern the point zero but also the exactness of a scale. This may happen for physical devices as well as for human observers. Other kinds of instrumentation are **floor** and **ceiling effects**.

Intention-to-treat analysis: In a **clinical study** a **random allocation** of the patients to the different treatment conditions is performed. Here, one also fixes the way in which the resulting data are recorded and analyzed. In an intention-to-treat analysis this schedule is observed in any case whether the patient has received the assigned treatment up to the end or whether this treatment was applied at all to the patient. Also compare **Zelen's single-consent design**.

Interaction: If the difference of the effects of two **levels** of an **independent variable** on a **dependent variable** depends on which levels of one or more other independent variables are present, an interaction is assumed.

Interaction effect: Common effect of two or more **independent variables** on a **dependent variable** which cannot be explained by the sum of the effects of the single independent variables. This effect is caused by an **interaction**.

Interindividual variation: Variation of the measurements for a group of subjects which is not exposed to any systematic effective exterior influences and where for each subject only one measurement is available.

Interinvestigatory affirmation: If **replications** are used and if the same **causal conclusions** can be drawn for many subjects or samples of subjects, the evidence for the found effects is increased and sometimes also the generalizability of the outcomes.

Interlaboratory trials: Studies used in order to check the **accuracy** of laboratory measurements. For this, test material is analyzed in laboratories which might differ with respect to location, staff, and used devices, and the laboratory outcomes are compared with each other.

Intermediary variable: Synonym for **intervening variable**.

Internal validity: The higher the internal validity of a study, the more **causal conclusions** are possible and the less it is possible to make observed effects implausible by means of **alternative explanations**.

Interrupted time-series design: Time-series design with **intervention**.

Interrupted time series with multiple replications: After a **baseline** measurement a treatment phase with **pretest** and **posttest** follows. After this a phase without treatment but with pretest and posttest is considered, then again a treatment phase with pretest and posttest etc. The interpretation of the outcomes can be improved if one randomly fixes, for each phase, whether a treatment is used or not (Cook and Campbell, 1979, pp. 222-223). Then, if in addition we have a **double-blind study**, it is even possible that **causal conclusions** can be drawn. For this, compare the **Edgington design**.

Interrupted time series with nonequivalent dependent variables: The **nonequivalent dependent variables design** is extended in that way that more than one pretest and more than one posttest is recorded for the two **dependent variables** (Cook and Campbell, 1979, pp. 218-221). **Causal conclusions** are not possible because, e.g., **maturation** might affect the two dependent variables in a different way.

Interrupted time series with a nonequivalent no-treatment control group time series: In two groups which were not formed by a **random assignment**, one **dependent variable** is measured simultaneously in both groups at different points of time. In one group an **intervention** is introduced but not in the other one (Cook and Campbell, 1979, pp. 214-218). Because of the absent **randomization**, **selection effects** cannot be ruled out, i.e. **causal conclusions** are not possible.

Interrupted time series with removed treatment: The **removed-treatment design with pretest and posttest** is extended in that way that before the treatment as well as during the treatment as well as after removing the treatment several successive recordings of the **dependent variable** are performed (Cook and Campbell, 1979, pp. 221-222). Because it is, e.g., possible that subjects react negatively to the removal of a treatment, **causal conclusions** are not possible.

Interrupted time series with switching replications: In two groups which have not been formed by a **random assignment**, a **dependent variable** is simultaneously measured during a long time interval. At a given point of time a treatment is introduced in one group, at another point of time in the other group (Cook and Campbell, 1979, pp. 223-225). At both points of time one group serves as a **treatment group**, the corresponding other one as a **control group**. Due to possible effects of **history**, causal conclusions are not possible.

Interval censored data: If the outbreak of a disease occurs for a patient between two routine checks the exact time of the outbreak is unknown and only a time interval can be given for which it is known that it contains the time of outbreak.

Intervening variable: If a **causal variable** has an effect on another variable which itself has an effect on an **effect variable**, the variable between causal variable and effect variable is called intervening variable. If two or more intervening variables are placed in a chain between the original causal variable and the effect variable, where always the preceding intervening variable is a causal variable for the succeeding intervening variable, this defines a **causal chain**. Obviously, each intervening variable in a causal chain is an effect variable for the preceding and a causal variable for the succeeding variable.

Intervention: A treatment which is applied in a **time-series design** at a given point of time or during a time interval.

Intervention study: A group of subjects is observed in a **longitudinal study** where no treatment is applied. At a fixed point of time the group is split up into two subgroups, one of which is furtheron without a treatment while the other one gets a treatment. After a fixed time period both groups are compared with each other to find out whether the **intervention** has had an effect.

Interviewer bias: If surveys are performed, interviewers are used to get answers from selected subjects. This can be one reason for biased outcomes. E.g., an interviewer might call on other subjects than those which he or she has been assigned to, or the questionnaires are completed by the interviewer, or the subjects are influenced by the interviewer in their answering behavior with or without knowledge of the interviewer.

Intraindividual variation: Fictive variation of the measurements of a subject which is not exposed to any systematic exterior influences. In case of **unobtrusive measures** the variation of the measurements in a **baseline** could be used as an example for intraindividual variation.

Intra-subject control: Synonym for **subjects as their own control**.

Intrinsic error: That variability in measurements which is due to the fact that each measuring device, at least as far as **quantitative responses** are recorded, shows a certain inaccuracy.

Inverse relationship: An increase of the values of one of two variables coincides with a decrease of the values of the other variable and vice versa.

Inverse sampling: Subjects of a **population** might exhibit a certain characteristic, e.g. a disease, or not. In case of inverse sampling, subjects are selected forming a **sample** up to that point of time, where a fixed number of subjects exhibiting the characteristic is achieved.

Irrelevant independent variable: Independent variable which has had no influence on a **dependent variable**. Because it is never possible to prove that there is no such influence, in reality no irrelevant independent variables exist.

Irreversible effect: An outlasting **effect** of an **independent variable**, which might be traceable, e.g., even after a **rest period**.

Isolated clinical case analysis design: By observing the behavior of single subjects one tries to formulate hypotheses about the causes for different behaviors. Because of the absence of **randomization** and other kinds of control, any obtained insights are purely speculative, and **causal conclusions** cannot be drawn.

Isomorphic block designs: Two **block designs** are called isomorphic block designs, if they differ only in the chosen order of treatments.

Item non-response: Questions in a **survey** which are not answered by the subjects. The relative frequency of such **missing observations** is given by the **non-response rate**.

ITT: Abbreviation for **intention-to-treat analysis**.

IV: Abbreviation for **independent variable**.

John Henry effect: Increased effort of subjects in a **control group** to attain the same performance as the subjects in a **treatment group**. This has the effect that actually the control group can no longer serve as a control group.

Joint effect: Synonym for **interaction effect**.

Judgement assignment: This is present in a **clinical trial** if either the doctor decides which treatment is assigned to which patient or if the patient decides which treatment he or she obtains. Then, one cannot rule out, e.g., that a certain treatment is assigned to patients which according to the opinion of the doctor or to their own opinion are seriously ill while other patients get another treatment. **Causal conclusions** cannot be drawn because **selection effects** cannot be ruled out.

Judgement sampling: A **sample** is selected from a **population** such that it is representative of the population according to the opinion of the selecting subject. See also **purposive sampling**.

Knight's move square: A **systematic design**, where the treatments are arranged by using the knight's move from chess. This is a generalization of the **Knut-Vik square**.

Knut-Vik square: A **Latin square** with five rows and five columns, where each of the five treatments occurs exactly five times. Here, equal treatments are connected by knight's moves as it is demonstrated in the following figure. According to Fisher (1966, p. 78) the design has been known in Denmark since about 1871 though it is usually ascribed to the Norwegian, Knut Vik.

A	B	C	D	E
D	E	A	B	C
B	C	D	E	A
E	A	B	C	D
C	D	E	A	B

Lagged dependent variable: This is the dependence of the value of a **dependent variable** at one point of time from values of the same dependent variable at preceding points of time, as it is to be assumed in **repeated-measures designs**.

Large simple trial: Clinical trial with a very large number of subjects, where the **inclusion** and **exclusion criteria** are defined such that they are not very restrictive. Only few and easy to measure **dependent variables** are recorded.

Last observation carried forward: If, in a **clinical study,** measurements at patients are not available up to the fixed point of time where the study is terminated, sometimes the last recorded measurement is substituted instead. This might cause wrong interpretations of the outcomes, in particular then, if the probability of a **missing value** depends on the treatment.

Latent factor: Synonym for **latent variable**.

Latent variable: Variable, which is used in a theory and which, as a rule, cannot be observed directly.

Latin cube: Generalization of a **Latin square** to control three **extraneous variables**. In the following example with two treatments A and B we give on the left the lower and on the right the upper layer of a Latin cube. It can be seen that each treatment occurs exactly once in each row and column of the depicted horizontal layers and that the same holds for the vertical layers.

AB	BA
BA	AB

Latin hypercube: Generalization of the **Latin cube** to more than three dimensions.

Latin square: For this kind of **incomplete counterbalancing** the number of experimental conditions, the number of points of time, and the number of considered sequences, i.e. also the number of groups, are identical. Each experimental condition occurs in each timely position and also in each group exactly once. In the following scheme with the four conditions A, B, C, and D the rows correspond to the different groups and the columns to the points of time. As a rule, **causal conclusions** are not possible as **asymmetric carry-over effects** cannot be ruled out. Latin squares are used not only in **repeated treatments designs** but also, if to each subject only one experimental condition is assigned and if two **extraneous variables** are to be controlled. In this case, the rows correspond to the **levels** of one extraneous variable and the columns to the levels of the corresponding other one.

A	D	C	B
D	A	B	C
C	B	D	A
B	C	A	D

Latin square crossover design: Orthogonal Latin squares can be used to generate **balanced crossover designs**, where the number of treatments is equal to the number of **periods**.

Lattice: Synonym for **lattice design**.

Lattice design: An **incomplete block design** which is a **proper design** and in which the **blocks** can be arranged in groups such that each group gets each treatment exactly once. The groups are called **replications**. In the following example 12 blocks, each with 3 subjects are assigned to 4 replications, where the replications are labeled with Roman numerals and the blocks with Arabic numerals over the block columns. Because the number of treatments (9) is equal to the square of the number of subjects for each block (3), this lattice design can also be considered as a **lattice square**. In this particular design not only the column blocks but also the row blocks are of importance, i.e. two **extraneous variables** can be considered at the same time. Because each pair of treatments occurs equally often (that is once) within the column blocks and within the row blocks of the example, we have a **balanced lattice square**. If one replication is omitted, a **partially balanced lattice square** remains. If the replications II and IV are omitted, a **semi-balanced lattice square** is generated, where each pair of treatments occurs once within a row or column block but not in both. If the schemes which correspond to the replications do not form squares, the designs are called **lattice rectangles**.

	I			II			III			IV		
	1	2	3	4	5	6	7	8	9	10	11	12
1	B	I	A	F	I	C	H	F	B	A	F	G
2	D	F	E	D	B	G	D	A	C	D	H	I
3	G	C	H	E	A	H	I	G	E	C	B	E

Lattice rectangle: See **lattice design**.

Lattice sampling: A geographical region is subdivided into equal-sized rectangles or squares, from which a **sample** is drawn. From the selected sub-areas all subjects, subsamples of subjects or single subjects are selected.

Lattice square: See **lattice design**.

Law of initial values: A pretended biological law detected by Wilder (1931). It predicts effects which have the same direction as **statistical regression**.

Lead time: The time interval between the detection of a disease for a patient participating in a **screening study** and the fictitious point of time where the illness would have been detected in a routine check or by perceiving corresponding symptoms.

Lead time bias: Estimates of the survival time are related to the time interval of the occurrence of symptoms for a disease to death. Therefore, it would be wrong to measure the survival time starting at the point of detecting the disease if still no symptoms are observable, e.g. if the disease is detected in a **screening study**.

Left-censored data: See **truncated sample**.

Left-hand truncated sample: Synonym for **left-sided truncated sample**.

Left-sided truncated sample: See **truncated sample**.

Length biased sampling: Diseases which need a long time to develop have a greater chance to be detected by medical routine checks than diseases which need only a short time to develop. As a consequence, e.g., slowly growing tumors are detected early in far more cases than rapidly growing tumors.

Level: Synonym for the value of an **independent variable**. A level corresponds to a specific experimental condition.

Life history method: Case study, where a subject is observed from the time of birth on. In most cases this is done while performing a **retrospective study**.

Limited data collection: This occurs if data are not recorded for single subjects. E.g., it might be suggestive, to pool the blood samples for a group of subjects to only one sample, in order to look for a disease agent. Only, if by means of this proceeding a disease agent is detected, the blood of the single subjects is tested.

Limited random sample: A **random sample** is selected from a **quota sample**, quite often even only from an **available sample**, because a **strict random sample** cannot be formed.

Line sampling: Synonym for **line transect method**.

Line transect method: A straight line is drawn across a geographical region and all subjects near to this line or on this line are selected to form a **sample**.

Literature control: A special case of historical control, where the data of the **control group** are taken from publications.

Local constancy: Term used for the **constancy** of known **extraneous variables** for the subjects of a **block**.

Local control: Control of extraneous variables separately for subsamples of subjects participating in an experiment. This is per-formed, e.g., by **matching** or **blocking**.

Local history: This is assumed in the case of an **interaction** between **selection** and **history**. If, e.g., **treatment** and **control group** are pre-existing groups which are not generated by a **random assignment**, one cannot rule out that these two groups are influenced by different events which operate from the exterior, because the two groups are, e.g., separated in space or time.

Local randomization: Random assignment of the subjects of a **block** to different experimental conditions.

LOCF: Abbreviation for **last observation carried forward**.

Longitudinal design: Synonym for **within-subjects design**.

Longitudinal study: Sometimes a **retrospective**, but in most cases a **prospective study**, where a **sample** of subjects is observed for a long time interval with respect to relevant characteristics.

Loss of individuals: Synonym for **experimental mortality**.

Lottery sampling: In order to be able to select a **random sample** from a **population**, a number is assigned to each subject and the random sample is selected from these numbers.

LST: Abbreviation for **large simple trial**.

Lurking variable: Variable which is responsible for an **illusory correlation** of two other variables.

Main effect: Distinct difference between measurements corresponding to the **levels** of an **independent variable** in a **factorial design**, if the average is taken over the measurements corresponding to the levels of all other independent variables.

Manifest variable: Observable variable which is often considered as a substitute for a non-observable **latent variable**.

Manipulated variable: Synonym for **independent variable**.

Marginal effect: Subjects at the edge of a **population** might respond in other ways to **experimental conditions** in comparison with subjects near to the centre of the population. If, e.g., according to **inclusion** and **exclusion criteria** only subjects between 20 and 40 years are admitted to a **clinical study**, different outcomes might be expected for 20 year old subjects in comparison with 30 year old ones.

Marginal matching: For this kind of **matching** it is not tried to form pairs of subjects which are similar with respect to given **matching variables**. Rather one tries to form groups of subjects which are similar with respect to certain characteristics of the considered matching variables, e.g., with respect to mean and variance of the matching variable age.

Masking: According to Mackintosh (1977, p. 491) masking is the effect that the effect of a target variable cannot be detected if several concurrent **independent variables** are effective at the same time. This occurs because the effect of the target variable is masked by the effects of other independent variables. The concept of masking should not be mixed up with the concepts of **covering** or **overshadowing**.

Master sample: A large **sample**, the so-called master sample, is selected from a **population**. From this sample subsamples are selected which are assigned to the **experimental conditions**.

Match by correlated criterion design: According to Matheson et al. (1971, pp. 47-48) an initial sample of subjects is subdivided by means of a **matching variable** into pairs of subjects which are as similar as possible. Within each pair one of two experimental conditions is randomly assigned to one of the two subjects (**randomization**), while the corresponding other subject gets the other condition.

Matched case-control study: A **retrospective study**, for which a **matching** is performed post hoc with respect to certain **matching variables**. Of course, no **randomization** and therefore no **causal conclusions** are possible.

Matched-groups design: See **matching by correlated criterion design**.

Matched pairs: The pairs of subjects which are used for **matching**. See **paired samples**.

Matched-pairs design: Synonym for **match by correlated criterion design**.

Matched set: See **one : *m* matching**.

Matching: This is a **local control** technique which is a special case of **blocking**. Pairs of subjects are formed such that given known **extraneous variables** have the same levels for the subjects of a pair, i.e., these levels are kept constant. The considered extraneous variables are also called **matching variables**. In case of matching two treatment conditions are considered, one of which is randomly selected (**randomization**) and assigned to the first subject of a pair. The corresponding other condition is assigned to the second subject. This proceeding is used for all pairs.

Matching variable: See **matching**.

Maturation: If a **repeated-measures design** is used, differences between a pretest and a posttest score corresponding to a treatment are not necessarily caused by an effect of the treatment but might be due to a change of the considered subject. Such changes which occur even though no effects of the experimental conditions are present are called maturation. A control of maturation is only possible if a **control group** without treatment is used.

Maturation effect: If subjects are interviewed at several successive points of time, differences might result just because the subjects become older. Compare also **age effect** and **maturation**.

Maximum response: In a **factorial experiment** this is the maximum value of the considered **dependent variable** which is recorded at a **factor level** (in case of one factor) or for a combination of factor levels (in case of two or more factors).

Maximum tolerated dose: The highest allowed dose of a drug with respect to toxicity. Compare also the **Fibonacci dose escalation scheme**.

MCAR: Abbreviation for **missing completely at random value**.

Measurement error: That portion of the observed variation of the measurements which is due to the inaccuracy of the measuring procedure. Compare also **error variance**.

Mediating variable: Synonym for **intervening variable**.

Medical audit: Recording of medical data in routine checks and treatments to get information where a quality improvement of the checks or treatments is possible.

Micro-data: See **aggregate data**.

Minimization method: This alternative to the **permuted blocks within strata** or the **biased coin method** was proposed by Pocock and Simon (1975). It is used to obtain approximately equal sample sizes in **clinical trials**. Assume the situation, where patients enter sequentially a study and have to be assigned immediately to a **treatment** or **control group**, and where in addition a **balancing** with respect to one or more **block factors** is scheduled. A patient enters the study at a certain point of time, and it is known for each combination of the **levels** of the block factors how many patients have already been assigned to each of the two experimental conditions. The new patient exhibits a certain level for each of the considered block factors. For each of these levels it is counted how many patients who have been assigned to the treatment group up to now exhibit this level. Then the sum of these counts is calculated. The corresponding sum is formed for the control group. If the two sums are equal the new patient is assigned with probability .5 to one of the two groups, as in situation 1 of the biased coin method. If the sum is larger for the treatment group, the patient is assigned to the control group with a probability of $p > .5$, e.g., $p = 2/3$, as in situation 2 of the biased coin method. If the sum is larger for the control group, the patient is assigned to the treatment group with probability p, as in situation 3 of the biased coin method. For more than two experimental conditions, the highest probabilities of assignment are given to the conditions with the smallest sums (Pocock and Simon, 1975; Pocock, 1979).

Missing completely at random value: **Missing value** for which the probability of occurrence depends neither on the magnitudes of the observed values nor on the magnitudes of the not observed values.

Missing observation: Observation which should be present according to the **study design** but which actually is missing, e.g., because a **censored sample** is present or because of **dropouts**.

Missing value: Synonym for **missing observation**.

Mixed classification: **Cross-classification** with at least three factors where at least for two factors a **complete cross-classification** is present but where also at least one nesting (see **nested design**) occurs.

Mixed design: Consider a design with two or more **independent variables**. If we have at the same time a **between-subjects design** with respect to some of the variables but a **within-subjects design** with respect to other variables, a so-called mixed design results.

Mixed effects model: Synonym for **mixed model**.

Mixed factorial experiment: Factorial experiment, where at least two **factors** have a different number of **levels**.

Mixed model: Here, for some of the **independent variables** the **levels** are determined according to the **fixed model**, while the levels for the other independent variables are determined according to the **random model**.

Mixed sampling: Here, several **sampling** techniques are combined in one sampling procedure. E.g., first one can try to achieve a **representative sample** by **quota sampling** and to select from this **random samples** in a second step.

Model experiment: Experiment, where the experimental conditions are generated artificially because they cannot be studied in reality for practical or ethical reasons. However, it is always an open question whether the results can be generalized to a real situation. E.g., it might be speculated that a certain memory component of human beings is influenced by a certain region of the brain. Because it is not allowed to destroy such a brain region in the subjects of an **experimental group**, the following ways of performing a model experiment might be conceived. First, one might try to give to the subjects in the experimental group an additional task for which the corresponding brain function is necessary thus blocking the use of this function for other tasks. Second, animals might be used instead of human beings, where a corresponding brain lesion in the experimental group might be tolerated.

Moderating effect: Synonym for **interaction effect**.

Moderator: Synonym for **moderator variable**.

Moderator variable: A variable, which influences the relationship between other variables. E.g., the observed quality of life might be high for patients in one hospital but low for patients in another hospital. If only patients are considered which receive no chemotherapy or only patients which receive chemotherapy, the observed quality of life is the same in both hospitals. In this case the variable "chemotherapy" with the two levels "present" and "absent" is a moderator variable which influences the relationship between the variables "hospital" and "magnitude of the observed quality of life".

Modified play-the-winner rule: See **play-the-winner rule**.

Mono-method bias: Even if a **construct** is represented by more than one kind of **operationalization** to avoid a **mono-operation bias**, one cannot rule out that all these operationalizations are realized in the same way, e.g., only by questionnaires but without a simultaneous recording of physiological measurements.

Mono-operation bias: If the **independent** and the **dependent variable** are represented only by one respective **operationalization**, one cannot rule out that existing **causal relations** cannot be detected because the operationalizations were not chosen in an appropriate way.

Monotonic decreasing trend: See **trend**.

Monotonic increasing trend: See **trend**.

Monotonic relation: This is present if in case of two variables an increase of the values of one variable is always accompanied by an increase (or always by a decrease, respectively) of the values of the other variable.

Mortality: Synonym for **experimental mortality**.

Most dissimilar case sampling: A most dissimilar case sampling is a **purposive sampling**, where **sample units** are selected which are as dissimilar as possible with respect to each other.

Most similar case sampling: A most similar case sampling is a **purposive sampling**, where **sample units** are selected which are as similar as possible with respect to each other.

MPW: Abbreviation for **modified play-the-winner rule**.

Multi-armed bandit allocation: Assume a slot machine with k arms. If one arm is pulled either the machine pays a unit or it pays nothing. The pay-off probabilities for the different arms are not known, might be different and are constant over time. For such a multi-armed bandit opti-mal game strategies are sought. The classical article about the two-armed bandit is due to Robbins (1952). Zelen (1969) pointed out that the problem is identical to the problem to assign patients who enter successively a **clinical trial** to the best of the k treatments. Actually, already Thompson (1933) studied the two-armed bandit and Thompson (1935) the multi-armed bandit with respect to this problem. Some possible strategies are the **play-the-winner rule**, and the **randomized play-the-winner rule** for the case with two treatments. By Bather (1980) the following strategy is studied for the multi-armed bandit: if a new patient enters the study, a randomized allocation index is calculated on the basis of the known outcomes for each of the k treatments up to the present point of time. Then, the treatment with the highest index is assigned to the patient. Such an index consists of two terms. The first term is the number of the former successes of a treatment divided by the number of the former applications of this treatment. The second term, which is always positive, is the outcome of a random experiment, where the influence of this outcome on the randomized allocation index is the smaller the more patients entered already the study.

Multicentre study: A **clinical trial** which is performed simultaneously at several hospitals or clinics on the basis of a **common protocol**. Often, for each centre a **random allocation** of its own is performed. It seems better to perform a common random allocation for all centres.

Multicentre trial: Synonym for **multicentre study**.

Multilevel design: Experimental design with an **independent variable** with more than two **levels**. It might be a **within-subjects design** or a **between-subjects design**.

Multioperationalization: The **operationalization** of a **construct** by several **manifest variables**.

Multiphase sampling: Generalization of **two-phase sampling**.

Multiphase screening: This is present if several checks are performed at the same point of time in a **screening study**, e.g., simultaneously a check with respect to breast cancer and a check with respect to intestinal cancer.

Multiple baseline design across behaviors: One assumes that several behaviors can be observed at a subject which are independent of each other. First, the frequencies of occurrence are recorded for all behaviors during a **baseline**. Then, one behavior is selected. In a treatment phase the same treatment is applied whenever this particular behavior is observed, while all other behaviors are, again, only recorded. In a next phase, the same treatment is applied additionally, if another selected behavior occurs etc. This means that the duration of a baseline is the larger, the later a behavior is considered in a treatment phase.

Multiple baseline design across settings: One assumes that a certain behavior of a subject can be observed in different independent settings. First, a baseline with respect to the frequency of occurrence of the behavior is recorded in all settings. Then, one setting is selected and a treatment is applied whenever the behavior is observed. In a next phase this treatment is also applied in another setting etc. Thus the duration of the baseline phase is the larger the later a setting is considered.

Multiple baseline design across subjects: In a block of subjects a certain behavior is recorded in a **baseline**. Then, a subject is selected and whenever the behavior is exhibited, a treatment is applied. After a fixed time interval the treatment is additionally applied to a second subject etc. The duration of the baseline phase is the larger the later a subject is considered. Also compare the **staggered baseline design**.

Multiple comparison design: Generalization of the **paired comparison design** to the case where in each comparison more than two subjects are to be compared.

Multiple endpoint: This is present if more than one **dependent variable** is recorded in a **clinical trial**.

Multiple schedule design: Only one behavior is observed at a subject. If this occurs, different treatments are applied in different phases. The different treatments are accompanied by different discriminating stimuli.

Multiple stratification: See **stratification**.

Multiple time response data: Points of time which are recorded for each afflicted patient in case of episodically occurring disease symptoms, as, e.g., in case of cardiac infarctions.

Multiple time-series design: Nonequivalent comparison group design, where at each subject not only one measurement but a sequence of measurements is recorded.

Multiple treatment design: Either an **experimental design** with one **independent variable** which exhibits more than two **levels** or an experimental design with more than one independent variable.

Multiple treatment interaction effect: Synonym for **carry-over effect**.

Multiplicative relation: Special form of an **interaction effect** which in this case is the product of the corresponding **main effects**.

Multi-stage experiment: In this kind of experiment it is not possible to realize the levels of an **independent variable** in one step but the levels can be realized only in several successive stages. E.g., the effect of noise on forgetting can only be studied, if learning has taken place at a preceding stage.

Multi-stage sampling: The selection of a **sample** is performed in several separate time phases, where always the same **dependent variable** is recorded, in contrast to **multiphase sampling**. Multi-stage sampling is an extension of **two-stage sampling**, i.e. from phase to phase always smaller units are sampled.

Multi-stage selection: See **selection**.

Natural experiment: A situation which is similar to an **experimental design**, e.g., because two groups can be considered as a **control** and a **treatment group**, but which is not caused by the intentional manipulation of an experimenter. Due to the absent **randomization** no **causal conclusions** can be drawn.

Natural-groups design: See **naturalistic observation study**.

Natural history study: Retrospective study, by which the typical course of a disease and possibly effective factors can be observed.

Naturalistic observation study: Observation of the behavior of subjects without that an experimenter manipulates any **independent variables**. No **causal conclusions** can be drawn.

Natural pairing: In case of animals from one litter or in case of twins, pairs of very similar subjects are obtained which are well suitable for **matching**. In the future an optimal natural pairing might become possible by cloning subjects.

Natural response: A response from the natural behavior set of a subject.

Natural setting: Situation which exists independent of whether a researcher wants to study it or not.

Naysaying: A response tendency of subjects where, independent of the context, subjects prefer to answer no instead of yes. Also see **yeasaying**.

***N*-by-*M* design: Factorial design**, where one **factor** exhibits *N* **levels** and the other factor *M* levels.

Negative relation: Synonym for **inverse relationship**.

Negative study: A study where no statistically significant result was obtained.

Neighbourhood controls: Synonym for **community controls**.

Nested case-control study: Subjects of a sample are followed up in a **retrospective study** until a certain disease breaks out. To each subject where the disease breaks out a suitable control subject from the sample is assigned for which the disease is not observed.

Nested design: A **factorial design** with at least two **factors**, for which an **incomplete cross-classification** is assumed. Further, a rank order of the factors is assumed such that each **level** of one factor can be combined only with one level of a factor with a higher rank. E.g., the factor A might exhibit the levels A1 and A2, the factor B the levels B1, B2, and B3, and the factor C the levels C1, C2, C3, C4, C5, and C6. In the following scheme, A has the highest, B the second highest, and C the lowest rank. E.g., C1 and C2 occur only together with B1 and A1, but not in combination with B2, B3, and A2. Therefore, C1 and C2 are nested under B1.

A1		A2			
B1		B2			B3
C1	C2	C3	C4	C5	C6

Nested sampling: Synonym for **multi-stage sampling**.

Nested variable: Variable which is an ingredient of a more comprehensive variable. E.g., the suicide rate for a town is nested under the suicide rate of the country where this town is located.

***n*-fold design:** A design, where at each place in an original 1-fold design a special treatment was scheduled, now the same treatment occurs *n* times. In the following example a 4-fold **Latin square** is given.

AA AA	BB BB
BB BB	AA AA

***n*-fold Latin square:** See ***n*-fold design**.

Nocebo: See **placebo effect**.

NOEL: Abbreviation for **no-observed-effect level**.

***N* of *1* clinical trial:** Synonym for **single-subject design**.

Noise factor: Synonym for **extraneous variable**.

Non-compliance: Insufficient **compliance**.

Non-current prospective study: Retrospective study where the existence of possible **risk factors** and the occurrence and the development of diseases of interest is concluded from available data bases.

Nonequivalent comparison group design: Two or more groups of subjects are formed on the basis of the values of certain **dependent variables**, i.e., not by **randomization**. **Causal conclusions** with respect to the outcomes of such a **quasiexperimental design** are not possible.

Nonequivalent control group design: Synonym for **nonequivalent comparison group design**.

Nonequivalent dependent variables design: Two **dependent variables** are measured at one group of subjects at the same point of time. Then a treatment is applied which is followed by a second simultaneous measurement of both variables. Here, it is supposed that the treatment might have at most an effect on the first dependent variable but not on the second one (Cook and Campbell, 1979, pp. 118-120). **Causal conclusions** are not possible, because, e.g., **maturation** might affect the two dependent variables in a different way.

Nonexperimental design: Synonym for **quasiexperimental design**.

Non-identified response: Censoring of an observation which is connected with the used **dependent variable**. An example is the refusal of patients to participate in a follow-up because they know that a certain measurement will cause them pain.

Non-informative censoring: In contrast to **informative censoring** one assumes that the probability for the occurrence of censored observations is independent of the experimental conditions. In particular, this probability should be independent from the point of time of the measurements. This kind of censoring is also called **random censoring** or **independent censoring**.

Non-masked study: Synonym for **open-label trial**.

Non-orthogonal design: Synonym for **unbalanced design**.

Non-randomized clinical trial: A **clinical trial** where, first, all participating patients get a new treatment. Those patients, where this treatment has a positive effect are further treated in this way, while the other patients get an alternative treatment, in most cases a standard treatment. Then, both groups of patients are compared with respect to one or more **dependent variables**. Because **selection effects** cannot be ruled out, **causal conclusions** cannot be drawn.

Non-random sample: Sample which was not obtained by **probability sampling**.

Non-responder: A subject which does not respond in the expected way to a treatment.

Non-response bias: See **non-response rate**.

Non-response rate: That proportion of the subjects in a **sample** for which a measurement of the **dependent variable** is not possible. Because one cannot rule out that the subpopulation of these subjects differs considerably from the total population it is not clear to which **population** any detected **causal conclusions** can be generalized. This is also denoted as **non-response bias**.

No-observed-effect level: That dose of a drug, below which a certain target response no longer can be observed.

No-treatment control group: A **control group**, where the researcher tries to avoid any kind of influence. By this kind of control group the detection of an **absolute effect** might be possible.

Novelty: Subjects who participate for the first time in a study might exhibit a behavior which differs from that behavior which they would exhibit outside the study. Reasons for this might be a certain shyness or the wish to present themselves as positive as possible in the sense of **social desirability responding**.

Null result: An empirical outcome which is not statistically significant. In such a case no statement about the existence or nonexistence of a **causal relation** can be made.

Number of replications: This number specifies how often a treatment is scheduled in an **experimental design**. These numbers might differ for different treatments.

Objectivity: A measuring method has a high objectivity, if the results do depend neither on the subject who performs the measurements, nor on the subject who evaluates the outcomes, nor on the subject who interprets the results.

Observational equality: This is given if a characteristic is recorded for all subjects in exactly the same way, i.e. also with the same precision.

Observational research: The researcher does not affect the observed subjects, neither by contact (e.g., an interview) nor by participating in the study.

Observational study: Synonym for **naturalistic observation study**.

Observation bias: Synonym for **information bias**.

Observed control approach: Treatment and control group are not formed by a random assignment of the conditions (**randomization**), but groups which already exist are used. E.g., a new vaccine is applied to children in one school and children from a school in the neighborhood serve as a control group. Due to possible **selection effects, causal conclusions** cannot be drawn from the outcomes of such a study.

Observer bias: Systematic errors in the recording of behavior by observers, which might be caused by the **Rosenthal effect**.

Observer drift: Deterioration of the quality of observers if these have no longer the impression that they are being controlled.

Observer variation: Systematic biases occur not only in case of subjective ratings, e.g., in the sense of **response bias** but also in case of seemingly far more reliable physiological measures. It is well-known, e.g., that there are doctors who systematically measure higher values of the blood pressure in the same patients than other doctors. In particular, in **multicentre studies** such biases can cause problems when the data are being interpreted. This becomes obvious, if **floor** or **ceiling** effects are present.

Occam's razor: See **Ockham's razor**.

Ockham's razor: Commonly used as a synonym for the **principle of parsimony**.

One group before-after design: In this one group design a treatment with pretest and posttest is scheduled. Due to the absence of a **control group** no **causal conclusions** are possible.

One group posttest only design: Synonym for **one shot case study**.

One group pretest-posttest design: Synonym for **one group before-after design**.

One : *m* matching: If **matching** is to be performed where a **randomization** is not possible, it sometimes happens that several subjects are available for each subject in the **treatment group** for a **control group**. In such a case, a fixed number ($m > 1$) of control subjects can be assigned to each treatment subject. The set of control subjects corresponding to a treatment subject is also denoted as a **matched set**.

One shot case study: This is the term by which Matheson et al. (1971, p. 34) denote a one group design, where only one treatment and a posttest is scheduled. Due to the absence of a **control group** no **causal conclusions** are possible.

One-step selection: Synonym for **single-stage selection**.

One-way classification: Experimental design with only one **factor**.

Open-label trial: If no **blinding** is scheduled in a **clinical trial**, i.e. if both the patient and the treating doctor know which treatment is being applied, we have an open-label trial.

Open protocol: Disclosure of the background conditions of an experiment for the participating subjects. This is indispensable, e.g., if in **clinical trials** an **informed consent** is required.

Open sequential design: See **sequential design**.

Open sequential sampling: See **sequential sampling**.

Operational definition: Hypotheses and theories are usually related to non-observable **constructs** or **latent variables**. This concerns **independent** as well as **dependent variables**. A test of such hypotheses or theories can only be performed if the latent variables have been replaced by observable **manifest variables**, i.e. if an operational definition of the latent variables has been considered.

Operational equality: Each **level** of a **factor** has to be realized in the same way for all subjects of a **sample**.

Operationalization: Use of a **manifest variable** instead of a **latent variable**. Compare **operational definition**.

Opinion survey: A **survey**, where the opinion of subjects from a given **population** with respect to certain topics is recorded.

Optimum response: Synonym for **maximum response**.

Option-3 scheme: In order to obtain stable measurement values, two measurements of the same characteristic are recorded and one then establishes whether their difference remains below a given threshold. If this is the case, the mean of the two values is used as a more stable substitute. Otherwise, a third measurement is recorded and the mean is calculated from those two values for which the difference has the smallest absolute value. If this proceeding is being used to obtain more stable pseudo-measurements, these values should not be entered into statistical calculations because problems might arise due to the resulting artificial reduction of variance.

Order effects: If several treatment conditions are applied to a subject in a **within-subjects design**, different effects might be observed for different arrangements of the conditions. A reason for this might be **asymmetric carry-over effects**.

Ordinal sampling: Synonym for **systematic sampling**.

Organismic variable: Specific **background variable**, which stems from the observed subject, e.g. sensitivity to pain.

Orthogonal Latin squares: Two **Latin squares** of the same order are called orthogonal Latin squares, if their combination yields a Greco-Latin square, i.e. if each letter combination occurs only once. In the following example first the two Latin squares are given and then the Greco-Latin square. Here, in the second Latin square A was replaced by α, B by β, and C by γ.

A	B	C
B	C	A
C	A	B

A	C	B
B	A	C
C	B	A

$A\alpha$	$B\gamma$	$C\beta$
$B\beta$	$C\alpha$	$A\gamma$
$C\gamma$	$A\beta$	$B\alpha$

Outcome variable: Synonym for **response variable**.

Outside influences: Changes in the environment of a subject between two times of measurement, which have not been caused intentionally by the researcher.

Overadjustment: This is present if in case of a **statistical adjustment** too much of the influence is being ascribed to **extraneous variables**.

Over-matching: An **independent variable** might influence a **dependent variable** not directly but in an indirect way via an intervening third variable. If this third variable is chosen as a **matching variable**, i.e. if it is kept constant within the **matched pairs**, no effect of the independent variable on the dependent variable is observed, though such an effect exists.

Oversampling: This is present if certain subpopulations occur with a higher percentage in a **sample** from a **population** than they would in the total population. If an oversampling occurs with respect to one or more subpopulations, this corresponds to an **undersampling**, i.e. to an underrepresentation, with respect to one or more of the remaining subpopulations.

Overshadowing: According to Mackintosh (1977, p. 491) overshadowing means that the presence of a strong stimulus might interfere with the effect of a weaker stimulus. The term overshadowing must not be mixed up with the terms **covering** or **masking**.

Overstratification: If too many **block factors** are used for a **blocking**, it might be difficult to find subjects for each combination of the **levels** of these block factors and it might be even more difficult to achieve the same sample size for each combination. This problem occurs, in particular, in **clinical trials** in which patients enter a study sequentially and where the method of **permuted blocks within strata** is applied.

Paired availability design: If a **random allocation** of patients in a **clinical trial** is not possible and if the influence of the **selection effect** is to be diminished, this design can be used if many pairs of **control** and **treatment groups** are at hand. Here, for all patients in the treatment groups a new kind of treatment is made available, even though not all of these patients might obtain this new treatment in reality. Similarly, the new kind of treatment is not available for the patients in the control groups, though some patients from the control groups might obtain the new treatment. Conclusions are drawn on the basis of the actually obtained treatment.

Paired Bernoulli data: The occurrence/ non-occurrence of two different characteristics is recorded for each subject. Here, four outcomes are possible for each subject: both characteristics are present, both characteristics are absent, the first but not the second characteristic is present, and, the second but not the first characteristic is present.

Paired comparison design: In a sample of subjects pairs are formed. For each of these pairs raters from a group of raters compare the two subjects.

Paired-groups design: Synonym for **match by correlated criterion design**.

Paired observations: Individual observations which are recorded for a pair of subjects with different treatments or for one subject after a pair of different treatments.

Paired samples: Samples, which were generated by **self-pairing**, **natural pairing** or **artificial pairing**.

Pairing: The formation of pairs which are as similar as possible when performing a **matching**.

Panel: See **panel study**.

Panel study: A group of subjects, the panel, is interviewed at several points of time with respect to certain topics. Compare also **wave**.

Parallel-dose design: Special **dose-ranging trial**, where to one group of subjects a **placebo** is given and to other groups different doses of a drug.

Parallel group design: Patients are randomly assigned to two or more arms, where to each arm corresponds another treatment.

Parallel individuals: Term for subjects from one **block**.

Parallel line assay: See **bioassay**.

Parsimony principle: Synonym for **principle of parsimony**.

Partial expectancy control: For **complete expectancy control** two **control groups** are used. If only one control group is used we have partial expectancy control.

Partially balanced change-over design: Synonym for **partially balanced crossover design**.

Partially balanced crossover design: Crossover design, in which each **period** contains all treatments equally often, but where the frequencies for sequels with two different treatments are not necessarily equal but might differ by one. In the following example the rows correspond to 3 succeeding periods, the columns to 4 subjects, and 4 treatments A, B, C, and D are considered. Sequels of two treatments occur either once or never.

A	B	C	D
B	A	D	C
C	D	A	B

Partially balanced lattice square: See **lattice design**.

Partial questionnaire design: In order to prevent low **response rates** which are caused by long questionnaires, all subjects of a sample have to answer the questions with respect to the **primary variables** but only a part of the sample also the questions with respect to the **secondary variables**.

Partial relation: Relation between variables which is observed within subsamples of a **sample** if a sample of subjects is subdivided in such a way into subsamples that a further variable is kept constant within the subsamples.

Participant observation: This is given if a researcher is a participant in the sample which is studied, whether the real participants are aware of this or not.

Partner studies: Subjects which are living together, i.e. under comparable conditions, are included in a study.

Passenger variable: A **dependent variable** which exhibits a relation with another dependent variable only because both variables are related with a third variable. Compare **illusory correlation**.

Passive variable: Causal variable which cannot be controlled by the researcher.

Patient refusal: See **patient withdrawal**.

Patient time: Time interval from entering to leaving of a patient in a **clinical trial**.

Patient withdrawal: Premature leaving of a **clinical trial** by a patient. This is a **protocol violation** which might yield the outcomes of a study uninterpretable if it occurs for many patients. Possible causes are either **patient refusal**, because patients refuse to participate furtheron in a study or **clinical judgement**, because patients are not to be treated furtheron in the scheduled way due to medical reasons, e.g., because complications are observed.

Peak value: Maximum response in a **dose-response curve**.

Percentage change: This results by dividing the **difference score** by the **pretest** score and by multiplying the ratio by 100. This is a special kind of a **gain score**.

Period: This is the point of time or the time interval where only one experimental condition is present in a **crossover design**. If the same condition is used more than once in a **sequence**, the corresponding number of periods is counted. E.g., in case of 2 experimental conditions A and B both sequences $A_1B_2A_3$ and $A_1A_2B_3$ have the same number of 3 periods.

Period effect: See **age effect.**

Periodical surveys: Surveys which are repeated in equal time intervals.

Permuted block design: A method of restricted **random allocation** in **clinical trials** to avoid that the sample sizes differ too much for different treatments. For this aim it is not advisable to assign randomly to each patient a treatment. Rather blocks of these patients are formed which enter the study by and by. It is assured that within each of these blocks each treatment occurs the same number of times. For this the number of patients within each of these blocks has to be a multiple of the number of treatments. In case of 3 treatments A, B, and C it is possible to assign, e.g., to always 6 successive patients randomly one of the 90 possible arrangements where exactly 2 patients correspond to each treatment. One of these arrangements would be, e.g., CAABCB. If the total number of patients is a multiple of 6, e.g. 30, all sample sizes are equal, e.g. 10. Otherwise, e.g. for the total numbers 31, 32, 33, 34 or 35, the sample sizes might differ at maximum by 2. As Efron (1971) remarked, it is a disadvantage of the permuted block design that after several treatment assignments within a block the experimenter knows with a high probability or even with certainty which condition is assigned to the next subject. To avoid this, Efron (1971) has proposed the **biased coin method**.

Permuted blocks within strata: If subjects enter a study successively and have to be assigned immediately to a **treatment** or **control group**, different proceedings are possible if in addition a **balancing** with respect to one or more **block factors** is to be performed. The worst way is the use of a

systematic design, where the subjects are assigned, e.g., in an alternating sequence to both groups, while taking into account the formation of blocks. Because of possible **selection effects** no **causal conclusions** can be drawn. However, in a **completely randomized design** one cannot rule out that the experimental conditions are not balanced within the blocks, i.e. that the sample sizes are unequal also within the blocks. One way to guarantee the occurrence of only small differences is the formation of permuted blocks within strata, i.e. for each combination of the **levels** of the block factors, i.e. for each **stratum**, an independent **permuted block design** is scheduled. As discussed under the entry permuted block design, the proceeding can easily be generalized to three or more experimental conditions. As Efron (1971) remarked, it is a disadvantage of the permuted block design that after several treatment assignments within a stratum the experimenter knows with a high probability or even with certainty which condition is assigned to the next subject. To avoid this, Efron (1971) has proposed the **biased coin method**.

Per protocol population: Subset of an intention-to-treat population which meets the following three conditions: 1. The subjects were exposed to the assigned treatment for a fixed minimum duration. 2. Measurements of the **primary variable** are available for relevant and fixed points of time. 3. No important violation of the **protocol** was observed, in particular, no violation of the **inclusion** and **exclusion criteria**.

Phase I study: After the effects of a new drug have been studied at animals, the first tests of the drug at healthy volunteers are performed in phase I. These tests are performed to evaluate possible risks and to study the metabolism and the bioavailability of the drug.

Phase II study: In this phase of a drug trial the first tests of a new drug at patients are performed, in order to find out the optimal dose with respect to effectivity and security.

Phase III study: This is a pharmacologic study with many patients, often a **multicentre study**, by which it is tested whether a new drug is superior in comparison with standard treatments with respect to security and effectivity.

Phase IV study: This is a **field study** which is performed over a long time interval to study the effects of a new drug after it had been released with respect to **side-effects**, morbidity and mortality.

Pilot experiment: Synonym for **pilot study**.

Pilot study: This is a study with only few subjects which is performed before the true experiment, to test whether it is actually possible to perform the main experiment in the way it is planned. Often it is argued that a pilot study is performed for testing, whether a main experiment should be performed at all after considering the outcomes of the pilot study. However, this is no serious argument because due to the small sample size no conclusions should be drawn from the outcomes of a pilot study. It should be made a difference between a pilot study and a **preliminary experiment**.

Pilot survey: Survey for a very small subpopulation, to assure the feasibility of the following actual survey.

Placebo: A treatment which cannot be discriminated from the true treatment by the subject but which differs from the true treatment in that respect that the effective component is absent.

Placebo control group: A control group where the experimenter uses a sham treatment. By this it is possible to prove the existence of **absolute effects**.

Placebo effect: It is often observed that patients who get a **placebo** exhibit a considerable improvement of their state of health in comparison with patients without any treatment and sometimes even in comparison with patients who got a true drug. This occurs though the placebo does not contain any known effective component. Sometimes it is also observed that after the application of a placebo the state of health deteriorates. In such cases the term placebo might be replaced by the term **nocebo**.

Placebo reactor: A patient which exhibits positive or negative **side-effects** after receiving a **placebo**, as it is usually only observed for effective substances.

Placebo response: Synonym for **placebo effect**.

Play-the-winner rule: A special **adaptive design** which was proposed by Zelen (1969). If two treatments are to be compared in a **clinical trial**, the treatment for the first patient is randomly selected. As soon as one knows that one of the treatments has had a positive effect, the probability for using this treatment is increased. If the treatments succeed much more rapidly than the outcomes of the treatments are known, the play-the-winner rule results in using both treatments approximately with the same probabilities. If the treatment outcomes are known before each succeeding treatment, this is called by Zelen (1969) the **modified play-the-winner rule**. For this rule, after each positive outcome the corresponding treatment is used furtheron, while for each negative outcome a switch to the corresponding other treatment is performed. An improvement of the play-the-winner rule and the modified play-the-winner rule is the **randomized play-the-winner rule** by Wei and Durham (1978). This rule avoids the disadvantages of the play-the-winner rule (known outcomes are not always taken appropriately into account) and of the modified play-the-winner rule (treatments are usually assigned to the patients in a deterministic way).

Plot: A plot or **whole-plot** is a given unit which restricts the freedom of choice within an experiment. In **split-plot designs** and **split-block designs** each whole-plot is subdivided into **sub-plots**.

Politz-Simmons technique: One problem arising in **surveys** is that many of the subjects which are interviewed are not met at home at the first contact. It is common practice to call on these subjects again and again until they are met at home. This proceeding causes high costs. Politz and Simmons (1949, 1950) developed a technique which requires only a single visit to the subjects. In addition to the interview those subjects which are met at home are asked at which time they are usually at home to determine that portion of time at which these subjects are at home during the hours which are scheduled for the interviews. The outcomes of the interviews are then weighted inversely to these time portions.

Population: The total set of subjects about which conclusions are to be drawn.

Positive relation: Synonym for **direct relationship**.

Post-sensitization: The term **sensitization** means that subjects respond more sensitively to a posttest if a pretest has been applied. If no pretest but a treatment has taken place, it is possible that the treatment has not only caused the expected treatment effect but also a sensitization, i.e. the subjects have become more sensitive to the posttest due to the treatment. Since treatment effects can only be measured by succeeding measurements of the **dependent variables** the true effect of a treatment and its sensitization effect cannot be isolated.

Poststratification: Here, in contrast to a **stratification**, a random sample of subjects is subdivided into **strata**.

Posttest: A measurement which is performed after an experimental condition has been introduced.

Posttest control group design: Synonym for **static group comparison design**.

Posttest-only design with nonequivalent groups: A **dependent variable** is measured only once in two groups which have not been formed by means of a **random assignment**. In one of the groups a treatment is applied before the measurement (Cook and Campbell, 1979, pp. 98-99). Because of the absent **randomization** no **causal conclusions** can be drawn. Also compare the **static group comparison design**.

Posttest-only design with predicted higher-order interactions: Two or more dependent variables are recorded in a group of subjects which is post hoc subdivided by means of a **covariate** into **strata**. Here, for the different strata different relations between the **dependent variables** are predicted on the basis of theoretical considerations (Cook and Campbell, 1979, pp. 134-136). Because of possible **selection effects** no **causal conclusions** can be drawn.

Potential independent variable: A variable which can be used as an **independent variable**, while its influence on a **dependent variable** still has not been studied.

Practice effect: One cannot rule out in **within-subjects designs** that a subject exhibits an increasing performance with increasing familiarity with the requirements of the experiment.

Pragmatic approach: Synonym for **intention-to-treat analysis**.

Pragmatic trial: A **clinical trial** with the object to acquire proposals for the treatment of future patients.

Precision: The precision of a method of measurement or of an experimental design is high, if the portion of not explained **error variance** is small.

Precision of a design: A design is the more precise, the smaller the portion of variation in the data which cannot be explained by the effect of the **independent variables**. Because the term precision of a design is not related to the possibility to draw **causal conclusions** but to statistical models for evaluating **designs**, irrespective of the violations of the assumptions of such models, one has to expect that the higher the precision of a design, the higher the number of possible **alternative explanations** and, thereby, the degree of non-interpretability.

Predictive research: In contrast to **exploratory research** the object here is to predict the values of a variable by the values of other variables without aiming at proving a causal relationship.

Predictor variable: Synonym for **explanatory variable**.

Preliminary experiment: In preliminary experiments one tests whether it is at all possible to realize a given **experimental design**. E.g., one tests by using only a few subjects whether it is possible to measure the used **dependent variables** with sufficient reliability and whether the **levels** of the **independent variables** were fixed such that existing effects can be detected. After the preliminary experiments have been performed one had better start a **pilot study** in order to check the experimental design as a whole with respect to its realizability.

Pre-post design: Synonym for **before-after design**.

Preselection: See **two-phase sampling**.

Pretest: A measurement which is performed before the experimental condition is introduced.

Pretesting: Synonym for **reactivity.**

Pretest-posttest control group design: Synonym for **before-after static group comparison design**.

Pretest-posttest design: Synonym for **one group before-after design**.

Pretest sensitizing: Synonym for **sensitization**.

Prevention trial: Clinical trial, which is used in order to examine the effect of treatments supposed to prevent the outbreak of a disease.

Primary end-point: Synonym for **primary variable**.

Primary survey: A **survey** to study a certain problem. If the recorded data are also used for the investigation of another problem, this is called a **secondary survey**.

Primary unit: See **two-stage sampling**.

Primary variable: This is that **dependent variable** which is, before a study is started, assumed to be the best suited for admitting **causal conclusions** with respect to the considered problem.

Primary variance: According to Matheson et al. (1971, p.18) that portion of the variation of the values of the **dependent variable** which is due to the **independent variable**.

Principle of parsimony: This principle requires that, if several explanations are at hand for an observed empirical relationship, the simplest one should be chosen. Actually, often the problem arises that it is not clear with respect to which aspect one explanation is simpler than another one.

Principle of testability: See **testability**.

Proactive history: By this are meant the individual properties as well as the individual learning history of subjects before starting a study. Differences of data from subjects with the same experimental condition are mainly ascribed to proactive history.

Probability sampling: For each possible **sample** from a **population** a probability of being selected is fixed. If this probability is the same for all samples of the same size, we have **simple random sampling**, otherwise **restricted random sampling**.

Prognostic factor: A **confounder** or a **covariate** in a **clinical study**. Also synonym for **explanatory variable**.

Prognostic variable: Synonym for **explanatory variable**.

Progressive error: According to Underwood and Shaughnessy (1975, p.65) these are possible effects in **within-subjects designs** which either cause a better performance of subjects with increasing experience (**practice effect**) or which cause a worse performance with increasing weariness, boredom or frustration.

Progressively censored data: When a **clinical trial** is being scheduled, the starting point and the end point of the study are usually fixed in advance. As a rule, not all patients are available at the beginning of the study, but they enter the study at different points of time. As a consequence, patients exhibit different sojourn times at the end of the study. Because measurements at patients can be recorded only up to the end of the study, the observation times and, thereby, also the censoring times might be different for different patients. As a rule, censoring times are the smaller, the later a patient has entered the study.

Prolective cohort: Feinstein (1973) introduced a new terminology to remove a certain confusion in the context of the meaning of **prospective study** and **retrospective study**. If it is studied, how birth weight is related to the weight of five-year-old children, children are followed up from birth to the age of five years. Thus, this **cohort** is followed up forward in time from cause to effect, i.e. we have a prospective study. However, if it is asked in which way the weight of five-year-old children can be explained by the birth weight, the weights of a cohort of five-year-old children is considered and these weights are connected in a retrospective study with the birth weights. Here, the cohort is followed up backward in time from effect to cause. Feinstein (1973) suggests to use here the term **trohoc** instead of the term "cohort", reading it backwards. If a study is scheduled such that at first the birth weights of a cohort of children are measured, we have a prolective cohort. However, if one starts with a cohort of five-year-old children for which the weights were recorded and if it is then tried to determine subsequently the birth weights of these children, we have a **retrolective cohort**. According to Feinstein (1973), the terms "prospective" and "retrospective" are related to the direction of conclusion (from cause to effect or vice versa), while the terms "prolective" and "retrolective" are related to the direction of data recording (forward or backward in time).

Proper design: A **block design** is called proper if each **block** contains the same number of subjects.

Prophylactic trial: Synonym for **prevention trial**.

Proportional allocation: A proportional allocation is present in a **stratified sample**, if the sample sizes for the different **strata** are proportional to the corresponding strata sizes. Here, also the term **proportional sampling** is used.

Proportional frequencies: In a **complete factorial experiment** a **sample** corresponds to each combination of the **factor levels**. If the following property holds for each **factor** we have proportional frequencies: Consider the sample sizes for the levels of a factor while the levels or combinations of levels for the remaining factors are kept constant. It should hold that while passing from one combination of levels for the remaining factors to another one, the sample sizes for the levels of the fixed factor have to be multiplied by a constant. If, in particular, this constant is always equal to one, the sample sizes for all samples are identical. In the following example the sample sizes are given for two factors with three or four levels, respectively. The second row results from the first one by multiplication with 2, the third row results from the first one by multiplication with .5. Similarly, the first column has to be multiplied by 2, 5 or 2, respectively, to obtain the second, third or fourth, respectively, column.

2	4	10	4
4	8	20	8
1	2	5	2

Proportional sampling: See **proportional allocation**.

Proportional stratified random sample: A **stratified random sample**, where the portion of subjects for each **stratum** is equal to the corresponding portion in the **population**.

Proportional subclass numbers: Synonym for **proportional frequencies**.

Prospective study: A sample of subjects is randomly split up into several subsamples (**randomization**) to which different treatment conditions are assigned. At the subjects measurements of **dependent variables** are recorded for a longer time, to be able to make statements about the different effects of the treatment conditions, after the study has come to an end. If such a study is performed as a **double-blind study**, **causal conclusions** are possible. The opposite of a prospective study is a **retrospective study**. For further discussion of the term prospective study see **prolective cohort**.

Protection: A control technique, where disturbing stimuli from the outside are taken into account by using, e.g., a blind or headphones.

Protocol: Document, where the whole proceeding for a **clinical trial** is fixed, in particular, the object of the study, the **inclusion** and **exclusion criteria**, the different treatment procedures, the study schedule, actions to be taken in case of **protocol violations** (e.g., in case of lacking **compliance** of patients) and the intended statistical evaluation.

Protocol violation: Each contravention of the **protocol** of a **clinical trial**, e.g., by **non-compliance** of patients.

Proxy measure: Synonym for **proxy variable**.

Proxy variable: Indirect measure of a variable to be studied which is used, if it is not possible to measure or observe the variable of interest in a direct way. This occurs, e.g., if a **manifest variable** is used instead of a **latent variable**. However, a proxy variable might also be an easy to measure manifest variable which is used instead of a difficult to measure manifest variable.

Pseudorandom number: A number which is generated by a mathematical algorithm in a deterministic way and which is used instead of a true **random number**, e.g. for drawing a **random sample**.

Purposive manipulation of the levels of the independent variable: The experimenter establishes on the basis of rational reflections which **levels** of the **independent variable** should be considered in his or her experiment and assigns to these levels samples of subjects in a random way (**randomization**). Using such a proceeding it is possible to draw **causal conclusions** in contrast to the design with **selection of the levels of the independent variable**.

Purposive sampling: Non-random selection of subjects from a **population** according to certain points of view.

PW: Abbreviation for **play-the-winner rule**.

Pygmalion effect: Synonym for **Rosenthal effect**.

Quadrat sampling: To get information about the composition of the **population** in a geographic region, quadrats are defined, i.e. areas of the same size and form. A fixed number of disjoint quadrats is randomly selected from the considered region. For each quadrat the exact composition of the contained sample of subjects is determined.

Qualitative factor: Factor, for which the **levels** are not defined by numbers or for which numbers serve only for naming the different levels but which otherwise might be chosen rather arbitrarily. An example are three different kinds of operation. However, if three doses of a drug are being considered, the assigned numbers have a real meaning and we have a **quantitative factor**.

Quantal assay: A series of increasing doses of a drug is considered and for each dose a **random sample** of subjects. For each subject it is observed, whether a certain effect occurred or not. In most cases one tries to estimate in this way the median effective dose (ED 50) or the median lethal dose (LD 50) for the drug.

Quantal response: A quantal response is observed, if a subject either exhibits a certain response after a treatment or if it does not. In contrast to this, a **quantitative response** informs also about the extent of the response.

Quantitative factor: See **qualitative factor**.

Quantitative response: See **quantal response**.

Quasiexperimental design: The decisive difference between a quasiexperimental and an **experimental design** is the absence of a **randomization** in the first design. Often also suitable **control groups** are missing in a quasiexperimental design. In general, no **causal conclusions** can be drawn on the basis of the outcomes of quasiexperimental designs.

Quitting ill effect: Often subjects renounce smoking as soon as disease symptoms caused by smoking are observed or if grave injuries to health are diagnosed. This has the effect that for the **population** of former smokers, who have in the meantime renounced smoking, a higher risk of lung cancer is observed than for the population of smokers who have not renounced. Similar effects are observed for alcoholics and drug addicts.

Quota sample: To derive a **representative sample** from a **population**, the population can be subdivided into **strata**, by means of suitable **extraneous variables**. This is done such that all subjects within a **stratum** exhibit the same combination of levels of the considered extraneous variables. As a rule, the strata will contain different numbers of subjects. A quota sample is a sample which exhibits the same composition with respect to the strata as the population. This means, in particular, that large strata have a higher representation in the sample than small strata. If possible, the subjects of a quota sample should be selected at random from the corresponding strata, Because here only a restricted **randomization** is effective, **selection effects** cannot be ruled out. I.e., the quota sample might be not representative with respect to certain extraneous variables which were not taken into account. In case of an unrestricted randomization a statistical representativeness of the **sample** for the population is achieved, whereby **causal conclusions** become possible. However, in such a **random sample** in contrast to a quota sample the strata might not be represented according to their size.

Random allocation: The experimental conditions are assigned to the subjects by an additional random experiment, e.g., by the toss of a coin (**randomization**).

Random assignment: Subjects are assigned to experimental conditions in a random way.

Random censoring: See **non-informative censoring**.

Random-effects model: Synonym for **random model**.

Random factor: A **factor** whose **levels** are selected randomly according to the **random model**.

Random-groups design: See **independent two group design**, though also more than two groups might be considered.

Randomization: Random assignment of subjects to experimental conditions or of experimental conditions to subjects. By this the effects of all known and unknown **extraneous variables** are controlled, at least from a statistical point of view, i.e. the probability of an erroneous **causal conclusion** can be kept below a fixed upper bound.

Randomization list: See **replacement randomization**.

Randomized block design: Blocks, i.e. subsamples of homogeneous subjects, are formed by means of **block variables**. Within each block subsamples of subjects are formed by **randomization**, and these subsamples are randomly assigned to the different experimental conditions. Thus, a randomized block design is a **complete block design** where besides **blocking** no other restrictions to randomization are present.

Randomized clinical trial: A **clinical trial**, where the patients are randomly assigned to the different treatments.

Randomized consent design: Two treatments A and B are to be compared in a **clinical trial**. After assuring the **eligibility** of a patient, he or she is randomly assigned to one of the two treatments. If a patient is assigned to treatment A, he or she is informed about the advantages and disadvantages of this treatment and is asked to consent to this treatment. If the patient agrees, he or she gets treatment A. Otherwise, an alternative treatment is applied, e.g., treatment B. If the patient is assigned by **randomization** to treatment B, he or she is asked in the same way to consent to this treatment after having been informed about treatment B. If the patient agrees, treatment B is applied, otherwise an alternative treatment. In this context compare also the items **intention-to-treat analysis** or **Zelen's single consent design**.

Randomized control trial: Clinical trial, in which the patients are assigned to the treatment conditions by a **random assignment**, and where at least one control condition, e.g., a standard treatment, is provided for. An improvement of this design is achieved by a **randomized double-blind placebo-controlled trial**.

Randomized design: Design, where the experimental conditions are randomly assigned to the subjects or where the subjects are randomly assigned to the experimental conditions (**randomization**).

Randomized double-blind placebo-controlled trial: Clinical trial, where the treatment conditions are randomly assigned to the patients, where a control condition with a

placebo is present, and where neither the patients nor the treating doctors know which treatment condition is applied. Because **causal conclusions** are possible for the outcomes of this design, it is nowadays a generally accepted standard design, for **phase II** studies.

Randomized matched posttest only control group design: Synonym for **match by correlated criterion design**.

Randomized pair: A pair of subjects which was generated by **pairing** and which is randomly split up to be assigned to two different experimental conditions.

Randomized play-the-winner rule: See **play-the-winner rule**.

Randomized posttest only control group design: Synonym for **randomized two group design**.

Randomized pretest-posttest control group design: Synonym for **before-after two group design**.

Randomized response technique: This is an application of the **dark room effect**. If there is reason to worry that subjects confronted with delicate questions are inclined to give answers influenced by **social desirability responding**, they get for each question a random generator and an instruction about the way to answer with the result that it is no longer possible to find out the true answer for a given subject even though this subject has not concealed the truth. E.g., subjects might be asked whether they ever had committed a shop-lifting. Together with this question they get a normal die and the following instruction: If the true answer is "yes" and one of the numbers 1, 2, 4 or 5 is observed when casting the die, give the answer "yes". However, in case of one of the two numbers 3 or 6 give the answer "no". If the true answer is "no" and one of the numbers 2, 3, 5 or 6 is observed, give the answer "no", in case of a 1 or 4 give the answer "yes". Because the researcher does not know the outcome of the casting of the die, he or she does not know for a given subject whether this subject answered according to truth or not. However, one assumes that the question is answered according to truth with a probability of 2/3 (i.e. in 4 out of 6 cases).

Randomized two group design: Synonym for **independent two group design**.

Random model: In this model the levels of the **independent variables** are assumed to be randomly selected from the set of all available levels. I.e. one does not regard these levels as arbitrarily fixed by the experimenter as this was the case in the **fixed model**. In contrast to the **selection of the levels of the independent variables**, in the random model subjects can be randomly assigned to the different randomly selected levels (**randomization**). Thus, **causal conclusions** are possible.

Random number: This is a number which is generated by a random mechanism, e.g., a die or a roulette wheel. The term is often used as a synonym for **pseudorandom number**, though this is not correct.

Random permuted blocks: Synonym for **permuted block design**.

Random permuted blocks within strata: Synonym for **permuted blocks within strata**.

Random sample: A random selection of subjects from a **population**.

Random sampling: Synonym for **simple random sampling**.

Random variation: Variation in the data which cannot be explained by any of the considered variables.

Ranked qualitative factor: A **qualitative factor** whose **levels** permit a natural ordering which, however, cannot be quantified. An example might be a factor "maze" with the three levels "very easy", "not easy but also not very difficult", and "very difficult".

Ranked set sampling: This is a **two-stage sampling** procedure proposed by McIntyre (1952). One assumes there exists a crude but not expensive method of measurement with respect to the considered characteristic. This allows an ordering of the subjects with respect to the characteristic but does not yield any measurement values. Further, an exact but expensive method of measurement might exist. From a **population** m **random samples** of subjects are selected, where each **sample** contains m subjects. At a first stage, the subjects are ordered within each sample by means of the crude method. At a second stage, from each sample a subject is selected on the basis of this ordering and at these m subjects an exact measurement is performed.

RBD: Abbreviation for **randomized block design**.

RCT: Abbreviation for **randomized clinical trial**.

Reactivity: Effect of the measurement of a **dependent variable** on subsequent measurements. Though many different effects might be conceivable, usually only **sensitization** and **resistibility** are mentioned. All these effects render it difficult to draw **causal conclusions**. This is why **repeated-measures designs** should be avoided.

Recall bias: In **retrospective studies** one cannot rule out that certain **risk factors** are reported far more often for the **cases** than for the controls. The reason for this is that for the cases a target causal variable is assumed which is absent for the controls.

Reciprocal relation: This is assumed if two variables are **cause** and **effect variable** for each other at the same time.

Recovery period: Synonym for **rest period**.

Reference population: If a **population** of ill people is given, the effect of a treatment is compared with the state of health of subjects in the population of healthy people. This latter population forms a reference population.

Regression artifact: Synonym for **statistical regression**.

Regression-discontinuity design: By means of a **covariate** a group of subjects is split up into two subgroups such that one group exhibits values of the covariate below a given value while the other group exhibits values of the covariate above this value. Now the subjects with the higher value of the covariate get a treatment, while the other subjects do not get this treatment. If the treatment has an **effect** on a **dependent variable**, the relationship between the covariate and the dependent variable should differ for the two subgroups (Cook and Campbell, 1979, pp. 137-146). **Causal conclusions** are not permitted, because the existence of such differences can only be proved, if very restrictive assumptions (e.g., a linear relationship) are made with respect to the kind of relation between covariate and dependent variable.

Regression effect: Synonym for **statistical regression**.

Regression to the mean: Synonym for **statistical regression**.

Reification: **Latent variables**, e.g., intelligence, memory, health or quality of life cannot be observed or measured in the real world. Further, **causal conclusions** can be drawn always only with respect to the **operationalizations** of the latent variables, i.e. with respect to the considered **manifest variables**. Nevertheless, it can be observed that very often latent variables are discussed as if they were really existing agents having an effect on dependent variables.

Related two group design: This might either be a **match by correlated criterion design** (Matheson et al., 1971, pp. 47-48) or a **before-match-after design** (Matheson et al., 1971, pp. 48-49) or a **yoked control group design**.

Relative effect: If one **treatment group** is compared with another treatment group in an **independent two group design** where the two groups get different treatments, an observed difference between the two groups with respect to the **dependent variables** is called relative effect. Also see **absolute effect**.

Relevant independent variable: Independent variable which has exhibited an effect on a **dependent variable**.

Reliability: Degree of exactness with which a measuring device measures a **manifest variable**.

Removed-treatment design with pretest and posttest: A treatment is applied to a group of subjects for a certain time. After this time the treatment is no longer applied. A **dependent variable** is measured just before the treatment period starts and at a point of time within the treatment period. A third measurement is performed at the beginning of the period with no treatment and a fourth measurement after a certain duration. The time intervals between the first and the second measurement on the one side and between the third and the fourth measurement on the other side should have the same length (Cook and Campbell, 1979, pp. 120-123). Here, we have an **untreated control group design with pretest and posttest**, where the two experimental conditions are applied to the same group of subjects in a fixed timely order. No **causal conclusions** can be drawn, because the subjects might exhibit a negative reaction to taking away an attractive treatment.

Repeatability: Similarity of measurements which are performed exactly in the same way, with only a short delay, with the same measuring devices, the same experimenter, in the same environment, and at the same subjects.

Repeated-measures design: If at each subject in an **experimental design** more than one measurement is performed, we have a repeated-measures design. Here, it is possible that before and/or after measurements are performed treat-ment and/or control conditions are scheduled. One cannot rule out that measurements have an effect on subsequent measurements (**reactivity**) or that **carry-over effects** of treatments occur. Therefore, in general no **causal conclusions** can be drawn from the outcomes of repeated-measures designs.

Repeated-measures factor: A **factor** for which the different **levels** are applied to the same subjects at different points of time using a **within-subjects design**.

Repeated sampling: In regular time intervals **samples** are selected from a **population**. In this way changes of the population might be detected.

Repeated treatments design: Within-subjects design, where the subjects are successively exposed to two or more treatment conditions.

Repeat examination: Repeated measurements of patients during the **run-in** in a **clinical trial** to check whether the **inclusion criteria** are really met, i.e. whether the patients exhibit really the diagnosed disease or whether they show only symptoms which disappear also without any therapy. An advantage of this proceeding is that subjects without the disease do not enter the study and are not exposed to a superfluous treatment. It might be wise to apply between two of the measurements in the run-in a **placebo**, to identify subjects who are not in need of a therapy, because a placebo might be sufficient to let disappear the symptoms.

Replacement randomization: If patients enter sequentially a **clinical trial**, it is possible to prepare before the study a **randomization list**. For generating this list it has been decided by means of an additional random experiment, e.g., by tossing a coin or casting a die, which treatment is applied to the first, second etc. patient. Of course, the randomization list should be prepared by a subject which cannot influence in any way the doctors or other subjects participating in the study. Rather, this subject is asked with respect to each patient entering the study which treatment should be applied. It is possible to prescribe to this subject which differences between the sample sizes might be tolerated at most. E.g., it might be required that for two treatments and a list with a maximum of 200 subjects both sample sizes might differ at most by 15. If this criterion is not met, the random experiment is repeated over and over again until a randomization list results which meets the requirement. Then, this list replaces all preceding lists which did not meet the criterion.

Replicated experiment: Exact repetition of an experiment. Sometimes this is also a **replication.**

Replicate observation: An observation which is repeated under the same conditions (as far as possible) as a first observation.

Replication: A replication is a repeated performance of an experiment, where the proceeding of the first experiment is exactly imitated. The term replication is also used with another meaning in the **lattice design**.

Representative equality: This is given if the outcomes for a **sample** can be generalized to the corresponding **population**.

Representativeness: This is given, if a **sample** from a **population** does not differ systematically in its composition from that of the population with respect to all known and unknown characteristics.

Representative sample: A **sample** from a **population** of subjects is called representative, if all subpopulations of the population are represented in the generally much smaller sample with the same portions as in the population.

Reproducibility: Similarity of measurements which are performed at the same subjects but with other measuring devices, other experimenters, and in other localities.

Resentful demoralization: A possible consequence of a **diffusion of treatments** can be that subjects of the **control group** respond embittered because they are disappointed that a certain treatment was denied to them. This might effect spurious differences between

treatment and **control group** which are either not due to the treatment at all or at least not to the observed extent.

Residual effect: Synonym for **carry-over effect**.

Resistibility: A form of the **reactivity** of measurements. A measurement of the **dependent variable** might have the effect that subjects respond to subsequent measurements less sensitive. This renders **causal conclusions** more difficult.

Respondent: Subject participating in an interview.

Responder: A subject that responds to a treatment in the way it was expected.

Response bias: Systematic error components which might alter the answers of subjects. For answers to questionnaires, e.g., a tendency is observed to tick the first or highest category (Mathews, 1927). Further, a **tendency to the extremes** (Hovland and Sherif, 1952) is observed, i.e. there is a preference for ticking the two extreme categories. An opposite tendency is the error of **central tendency** (Guilford, 1954, pp. 278-279), where middle or neutral responses are preferred. In Guilford (1954, p. 278) the **error of leniency** is described according to which subjects which are better known to an observer are judged more lenient than subjects which are less well known. The opposite effect is also reported. A further tendency of this kind is the **halo effect** which was described by Thorndike (1920). Lentz (1938) describes **acquiescence** as the **tendency to agree**.

Response rate: Relative portion of subjects responding to an interview.

Response reactivity: Synonym for **reactivity**.

Response surface: If a **quantitative response** is predicted by means of two or more **quantitative factors**, the corresponding geometric visualization is called response surface.

Response variable: Synonym for **dependent variable**.

Responsivity: Synonym for **reactivity**.

Rest period: Periods between the treatments in **crossover designs**, where either no treatment or a standard treatment is given. By rest periods one tries to avoid **carry-over effects** or at least to diminish them.

Restricted randomization: The random assignment of the experimental conditions to the subjects (**randomization**) is restricted by other control procedures, e.g., by **blocking**. This is in contrast to a **completely randomized design**.

Restricted random sampling: See **probability sampling**.

Retroactive history: Synonym for **history**.

Retrolective cohort: See **prolective cohort**.

Retrospective cohort study: In a **retrospective study** a **cohort** is identified, its case history is studied and compared with the case history of a suitable **control group**.

Retrospective study: For a sample of subjects it might be known which state of health (e.g., dead or alive, healthy or ill) the subjects exhibited at a certain point of time. Further, measurements for one or more **dependent variables** might be available for these subjects for a long time interval before the point of time where the state of health was established. Now, one tries to formulate subsequently hypotheses about possible causes for the observed state of health. Due to absent **randomization** no **causal conclusions** are possible. The opposite of a retrospective study is a **prospective study**. See also **prolective cohort**.

Retrospective survey: Synonym for **retrospective study**.

Reversal design: Two incompatible kinds of behavior B_1 and B_2 are observed at subjects. First, for each kind of behavior a **baseline** with respect to its frequency of occurrence is recorded. Then, after each behavior B_1 a treatment T_1 is applied and after each behavior B_2 a treatment T_2. After a certain time interval behavior B_1 is followed by treatment T_2 and behavior B_2 by treatment T_1, i.e. there is a reversal of the treatments. Often a new reversal follows after a certain time, i.e. the original sequence of treatments is reestablished. By such a reversal design which should not be mixed up with a **withdrawal** or **ABA design**, the relative effectiveness of the treatments T_1 and T_2 is to be proved. As it is true for most **within-subjects designs** no **causal conclusions** are possible because the effects of the different treatments cannot be isolated as a rule.

Reversed-treatment nonequivalent control group design with pretest and posttest: In two groups which were not generated by a **random assignment**, a **dependent variable** is measured twice, where in one group a treatment is applied between the two measurements, while in the other group another treatment is applied between the two measurements. For the two treatments an opposite effect is expected (Cook and Campbell, 1979, pp. 124-126) . Because of the absent **randomization** one cannot rule out that **selection effects** exist. Thus, **causal conclusions** are not possible.

Right-censored data: See **truncated sample**.

Right-hand truncated sample: Synonym for **right-sided truncated sample**.

Right-sided truncated sample: See **truncated sample**.

Risk factor: A potential **causal variable** for which it is either known that it has an effect on the genesis or the course of a disease or for which it is known that there exist corresponding correlative relationships or for which it is known that it has an effect on other causal variables which in turn have a corresponding effect with respect to the disease.

Risk set: The set of those patients in a **clinical trial** who a short time before a fixed point of time are neither dead nor censored.

Rosenthal effect: This denotes effects on a **dependent variable** which are caused by expectancies of the experimenter. Thus, such expectancies might be effective as **extraneous** variables. This is a specific **experimenter effect** or **experimenter bias**.

Rotatable design: Symmetric design for getting data for estimating a **response surface**.

Rotation sampling: A **repeated sampling** is used, where at each point of time only a part of the subjects is newly selected, while at the remaining subjects a second measurement is recorded.

Round robin design: A **paired comparison design**, where each pair of subjects is judged the same number of times.

RPW: Abbreviation for **randomized play-the-winner rule**.

Run-in: To diminish the probability of **dropouts** which are due to **non-compliance** of patients in **clinical trials**, an observation phase or run-in is used before the **random allocation**, where the behavior of the patients is judged with respect to the **protocol**. Patients with unsatisfactory compliance are excluded from the study before the random allocation is performed.

Sample: A sample is a selection of subjects from a **population**.

Sample size: Number of patients in a **sample**.

Sample survey: Collection of data from a systematically selected subpopulation.

Sample unit: In most cases **samples** are formed by selecting subjects from a **population**. Then the subjects are the sample units. However, in case of **cluster sampling** groups of subjects are the sample units, at least at the first stage, and in case of **area sampling** the sample units consist of sub-areas of a region.

Sampling: The selection of **sample units** from a **population**.

Sampling bias: The **population** which is available for selecting a **sample**, often has a composition which is totally different from that of the **target population**. Therefore, relationships which were found for the samples often cannot be generalized to the target population. Thus, the relative proportions of diseases for patients in hospitals are often very different from the corresponding proportions in the population due to the different modalities of the hospitals with respect to the acceptance of patients. Another kind of sampling bias occurs if the more compliant patients remain in the study until it ends, while the not-so-compliant patients leave the study at a former point of time.

Sampling error: Synonym for **sampling bias**.

Sampling fraction: Ratio of **sample size** and population size.

Sampling frame: Synonym for **frame**.

Sampling variation: If several **samples** of the same size are selected from a **population**, often a different **interindividual variation** is observed in the samples though no **independent variable** was varied.

Sampling with arbitrary probability: Before **sampling** it is fixed with which probabilities single subjects might enter a sample. E.g., in **cluster sampling** at a first stage households might be selected and only at a second stage subjects from households. Depending on the problem it might be better to select households with many subjects with a higher probability than households with few subjects.

Sampling with equal probability: Each **sample unit** has the same probability to be selected from a **population**. This might be problematic, in particular, in case of **two-stage sampling**, if the sample units at the first stage have different sizes. Compare **sampling with arbitrary probability**.

Sampling without replacement: A **sample** is selected from a **population** by removing one **sample unit** after the other without returning it to the population. In case of a finite population, in general, each selection of a sample alters the composition of the population. In contrast to **sampling with replacement** all sample units in a sample are different.

Sampling with replacement: A sample is selected from a **population** by selecting one **sample unit** after the other, where each sample unit returns to the population immediately after it was selected. The selection of the sample does not change the composition of the population, even if the population contains only a finite number of units. In contrast to **sampling without replacement**, one or more sample units might occur more than once in the sample.

Screened-to-eligible ratio: In a **clinical trial** the **eligibility** of patients who might enter the study is fixed by the **inclusion** and **exclusion criteria**. As a rule, at the start far more patients than necessary have to be considered, to obtain the number of eligible subjects which is scheduled for the study. The screened-to-eligible ratio is here the ratio of the number of patients who actually entered the study and of the number of patients who had to be considered for the selection procedure.

Screening: Multistage selection procedure to select a certain subpopulation, e.g., ill people, from a **population**. Here, at a first stage a rather crude but low-priced procedure is used. At a second stage a more efficient though also more expensive procedure is applied to those subjects who were not selected for the subpopulation at the first stage. In a similar way a third and more stages can be conceived. Such a screening can also be used to lower the **non-response rate**.

Screening study: Study, to detect in a **population** subjects with certain diseases at a very early stadium. In a **continuous screen design** subjects are randomly assigned (**randomization**) either to a group which is examined in regular time intervals or to a group which is not examined in this way. Less expensive is the **stop screen design**, where the regular examination of the one group is only performed for a fixed time interval.

Secondary survey: See **primary survey**.

Secondary unit: See **two-stage sampling**.

Secondary variable: Supplementary variable which is recorded for facilitating the interpretation of the outcomes for a **primary variable**. Sometimes also used as a synonym for **extraneous variable**.

Secondary variance: According to Matheson et al. (1971, p. 18) that part of the variation of the values of the **dependent variable** which can be explained by the effects of known **extraneous variables**.

Second order correlation: Correlation, i.e. a linear relationship, between two variables, while two other variables are kept constant. This might be considered, e.g., for avoiding an **illusory correlation**.

Second order design: Design for attaining with three **levels** for each **factor** the data for estimating a **response surface**.

Second order rotatable design: Symmetric **second order design**.

Second-stage unit: See **two-stage sampling**.

Selection: Selection or **artificial selection** means that subjects are selected from a **population** on the basis of the values of certain variables. If this selection is performed with respect to the values of one variable, we have a **direct selection** with respect to this variable and an **indirect selection** with respect to other variables. If several variables are considered simultaneously in the selection, we have a **combined selection**. If the selection is performed at one stage we have a **single-stage selection**, otherwise a **multi-stage selection** or **sequential selection**. If at the different stages of a multi-stage selection different information is used with respect to the considered variable, this is a **tandem selection**.

Selection bias: Synonym for **selection effect.**

Selection effect: If in an **experimental design** subjects are not assigned by chance to the experimental conditions or if alternatively experimental conditions are not by chance assigned to subjects, as it would be the case for an appropriate **randomization**, then selection effects cannot be ruled out. By this is meant that systematic differences between the measurements of the **dependent variable** for different experimental conditions might be partly or totally due to pre-existing systematic differences between subjects of different groups and are not necessarily caused by the differences between the experimental conditions. Therefore, **causal conclusions** are not possible.

Selection of the levels of the independent variable: On the basis of the levels of a **dependent variable** subjects are assigned to groups and the original dependent variable is now interpreted as an **independent variable.** This means that the **levels** of the independent variable are not assigned by the experimenter in an arbitrary way to the subjects, but this assignment exists independently of the experimenter. In contrast to the **purposive manipulation of the levels of the independent variable** no **causal conclusions** are permitted because a **randomization** is not possible.

Selection threat: Synonym for **selection effect**.

Self-conjugate Latin square: A **conjugate Latin square** which is conjugate to itself, i.e. each row is equal to the corresponding column. From this follows that a self-conjugate Latin square has to be symmetric with respect to the main diagonal as it is also the case in the following example.

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

Self-pairing: One tries to pair a subject with itself by applying two experimental conditions to the same subject at two different points of time.

Self-selection bias: Selection effect, which occurs if subjects are not randomly assigned to the experimental conditions but assign themselves to these conditions.

Self-weighted sample: A self-weighted sample is given if after the selection of a **stratified sample** the **samples** from the different **strata** have **sample sizes** which are proportional to the corresponding strata sizes.

Semi-balanced lattice square: See **lattice design**.

Sensitization: A form of the **reactivity** of measurements. A measurement of the **dependent variable** can effect that subjects react more sensitive to subsequent measurements. This renders **causal conclusions** more difficult.

Sequence: By this is denoted the timely order of experimental conditions in a **crossover design**.

Sequential design: For a **clinical trial** we have a sequential design if the **sample size** is not fixed before the trial. Rather, each time when the data for a patient or for a fixed number of patients in a **group sequential design**, respectively, are available, it is decided whether already a **causal conclusion** can be drawn or whether more patients have to be recruited for the study. In case of an **open sequential design** it is, at least in theory, possible that such a study never terminates. In **closed sequential designs** the end of the study is defined, e.g., by a bound for the maximum total number of participants.

Sequential sampling: Samples of subjects are selected in subsequent time intervals, i.e. no **fixed sample size** is assumed. Each sample consists either of one subject or of several subjects. After each time interval it is decided on the basis of the observed outcomes whether the considered question can be answered or if further samples have to be selected. If no restriction is formulated with respect to the number of samples which can be selected we have an **open sequential sampling**. However, if the maximum total number of subjects which is to be selected is fixed we have a **closed sequential sampling**.

Sequential selection: See **selection**.

Serendipity: According to McGuigan (1978, pp. 64-66) the perception of the possible importance and the subsequent investigation of strange and unexpected observations during an experiment which seem to be or are in fact not connected with the **experimental design**.

Serial measurements: Measurements at the same subject at subsequent points of time.

Series of experiments: In experiments quite often the **levels** for many **extraneous variables** are kept fixed to facilitate **causal conclusions**. Such variables are also called **constant factors**. In a series of experiments the same experiment is performed for different combinations of the levels of the constant factors. Thereby the generalizability of the outcomes as expressed by **external validity** is improved.

Setting: The situation arranged by the experimenter where an experiment is performed.

Sham operation: In experiments on animals an operation, where the operation corresponds to the corresponding one in the **treatment condition** in all but one respect. The exception concerns the really decisive step in the intervention. By this control condition the nonspecific effect of the operation is to be controlled.

Side-effect: Effect of an **independent variable** on **dependent variables** for which no effect is to be investigated. In many cases side-effects are unwelcome effects of treatments (**adverse effects**).

Simple classification: Synonym for **one-way classification**.

Simple comparison: See **complex comparison**.

Simple effect: Distinct differences between the measurements for the **levels** of an **independent variable** in a **factorial design**, if the levels of all other independent variables are kept fixed.

Simple experimental design: Either a synonym for **completely randomized design** or for a design with only one **factor**.

Simple interrupted time-series design: Synonym for **time-series design** with **intervention**.

Simple random sampling: See **probability sampling**.

Single blindfold experiment: See **blinding**.

Single-blind study: In such studies the participating subjects are not able to distinguish the different experimental conditions. Also compare **blinding**.

Single-case experimental design: Synonym for **single-subject design**.

Single experiment: An experiment, where it is possible to keep the **levels** of a set of **extraneous variables** constant for the duration of the experiment.

Single-masked study: Synonym for **single-blind study**.

Single-participant design: Synonym for **single-subject design**.

Single replication design: A design, where each experimental condition occurs exactly once.

Single-stage selection: See **selection**.

Single-subject design: **Within-subjects design** with a sample consisting of only one subject. In most cases no **causal conclusions** are possible. An exception are the **Edgington designs**.

Singly censored data: If a **clinical trial** starts at a fixed point of time and terminates at a fixed point of time and if all patients enter the study at the same point of time, we have singly censored data, if all relevant measurements or observations, respectively, can be recorded at all patients until the end of the study, i.e. if no patient is lost due to **protocol violations**. The time point of censoring is at the same time the time point of termination of the study. If patients enter the study at different points of time but it is guaranteed that each patient is observed for a fixed time interval and if this interval is not shortened or increased by fixing the point of terminating the study, we have again singly censored data. However, if the time points of starting and terminating the study are fixed and if the patients enter the study at different points of time, **progressively censored data** result. This means that patients who enter the study at an earlier point of time have participated a longer time in the study at the time of termination than patients entering the study at a later point of time.

Six-point assay: Specific six-group design in pharmacologic studies to compare a standard drug with a new drug. For each of the two drugs three doses are considered. The difference (or the ratio, respectively) of the third dose to the second dose should be the same as the difference (or the ratio, respectively) of the second dose to the first one.

Sleeper effect: An effect which follows with a certain time delay after a cause.

Slope ratio assay: See **bioassay**.

Snowball sampling: A **sample** is selected from a **population** of subjects. Each subject is asked to name further subjects who might participate in the study. This proceeding can be continued as long as new subjects are still being named. This method might be used if one wants to draw samples from subpopulations which are difficult to approach, e.g., homosexuals or drug addicts.

Social desirability bias: Synonym for **social desirability responding**.

Social desirability responding: Subjects try to appear as positive as possible to the experimenter. This is a kind of **impression management**.

Solomon four group design: This **experimental design** is used in order to study the **reactivity** of pretests. For this a **factorial design** with two **factors** is considered. One factor has the two **levels** "treatment (T)" and "control (C)". The other factor has the two levels "pretest (P)" and "no pretest (N)". This yields the four factor level combinations P-T, P-C, N-T, and N-C. A sample of subjects is randomly split up into four subsamples corresponding to the four combinations (**randomization**). After the application of each combination the outcome of a posttest is recorded. By comparing the posttest outcomes after P-T and N-T or after P-C and N-C, respectively, it can be established whether the pretest has had an effect on the posttest. By a comparison of the two differences one can conclude whether an **interaction** between pretest and treatment exists.

Solomon three group design: If the factor level combination N-C is not considered in the **Solomon four group design**, the Solomon three group design results.

Split-block design: A **complete cross-classification** is applied to **blocks**. For this, a plot is first subdivided with respect to the levels of one factor and this subdivision is then crossed with respect to the levels of the other factor. E.g., a market research institute might have eight test households in each of three towns. Always two of the households are in the same house, one on the lowest floor, the other one on one of the highest floors. The institute gets the order to test two different types (W1 and W2) of washing-machines as well as two different types (D1 and D2) of dryers. For this, in each of the three towns each of four households from two houses gets a washing-machine of the type W1 and each of the remaining twelve households a washing-machine of the type W2. In addition, in each of the three towns the four households on the lowest floors get a dryer of the type D1, the remaining households get a dryer of the type D2. In this way, in each of the three towns each of the four possible combinations of a washing-machine with a dryer is assigned to two households. In this example, each of the three plots (towns) is first subdivided with respect to the levels W1 and W2 of factor W. This subdivision is then crossed with the levels D1 and D2 of factor D.

Split-litter technique: A special **blocking** technique, where the animals of one litter are randomly split up and assigned to the treatment conditions (**randomization**).

Split-plot design: In a split-plot design at first a **plot** is randomly assigned (**randomization**) to each **level** of one **factor**. Each of these plots or **whole-plots** is then subdivided into as many smaller **sub-plots** as a second factor has levels. To each sub-plot of such a whole-plot a level of the second factor is randomly assigned. E.g., a market research institute might have in each of three towns twelve test households, where always exactly two of the households are in the same house. The institute gets the order to test two different types (W1 and W2) of washing-machines as well as two different types (D1 and D2) of dryers. For this, at first three houses (the whole-plots) in each of the three towns are randomly selected and each of the altogether 18 households in these houses gets a washing-machine of type W1. Each of the remaining 18 households gets a washing-machine of type W2. Then, in each house one of the two test households (the sub-plot) is randomly selected, and to this household a dryer of type D1 is assigned. To the other households a dryer of type D2 is assigned. From this example it can be seen that a split-plot design is no **completely randomized design** but a design with **restricted randomization** because in each house the washing-machines are of the same type and in no house are dryers of the same type. By considering a third and further factors the split-plot design can be extended by further subdividing the sub-plots.

Spurious correlation: Often used as a synonym for **illusory correlation**. A more restricted meaning of spurious correlation concerns a relationship which is observed in the data and which is caused by a **sampling bias**.

Spurious relation: Synonym for **spurious correlation**.

Staggered baseline design: According to Robinson (1976, pp. 273-274) this is a design where at several subjects first a **baseline** is recorded and then a treatment is applied. The point of time, where the treatment is applied is varied between the subjects. By this kind of proceeding one tries to diminish the probability that effects occur only for that reason that by chance at the starting point of the treatment an **extraneous variable** is effective at the same time. With this design **causal conclusions** are not possible. Also compare the **multiple baseline design across subjects**.

Standard design: Synonym for **Fibonacci dose escalation scheme**.

Standard gamble: Synonym for **Von Neumann-Morgenstern standard gamble**.

Standardization of treatments: Exact fixing of the conditions in an experiment so that it can be performed by different experimenters in exactly the same way.

Standard Latin square: A **Latin square** in which the treatments are listed in the first row and in the first column in alphabetic order as in the following example.

A	B	C
B	C	A
C	A	B

Static group comparison design: According to Matheson et al. (1971, pp. 42-43) an **ex post facto design** or a **quasiexperimental design**, respectively. Here, one sample of subjects gets a treatment, another sample gets no treatment. The subjects are not randomly assigned to the two treatment conditions, i.e. no **randomization** takes place. Because **selection effects** cannot be ruled out, **causal conclusions** are not possible.

Static population: See **dynamic population**.

Statistical adjustment: The attempt to isolate the influence of **extraneous variables** by means of statistical procedures, e.g., by means of an analysis of covariance.

Statistical conclusion validity: This is the higher, the more it is justified to conclude the existence of genuine effects on the basis of empirical data.

Statistical regression: A statistical artifact which might result in low posttest values after high pretest values and in high posttest values after low pretest values. The direction of the effect of statistical regression is the same as that of the **law of initial values**. The effect has to be expected, e.g., if **extreme groups** are formed with respect to a variable. E.g., prior to a course a group of students is subdivided into groups of students with high, normal, and low scores on the basis of a psychological test. For the corresponding measurements after the course it might be found that the "good" students exhibit lower scores than before, while the "bad" students exhibit higher scores.

Statistical twins: Expression for pairs of subjects which were formed by means of **pairing**.

Stimulus-response specifity: If a nonobservable **latent variable** is operationalized by an observable **manifest variable** and is used as an independent variable whose effect on a **dependent variable** can be shown, we have an interpretational problem if the used manifest variable might also be considered as an operationalization of other latent variables. In this case it is not clear which of the possible latent variables is responsible for the observed effect.

Stop screen design: See **screening study**.

Strata: Plural of **stratum**.

Stratification: Dissection of a **population** into nonoverlapping subpopulations which are denoted as strata. The strata should be as homogeneous as possible with respect to given characteristics. By selecting a **random sample** from each **stratum** one tries to obtain a **representative sample**. The result is a **stratified sample**. If stratification is used not only with respect to one characteristic, we have a **multiple stratification**. If the number of subjects in substrata is known for a multiple stratification, this is also called a **control of substrata**.

Stratified assignment: Synonym for **blocking**.

Stratified nonrandom sample: Quota sample, which is not formed by **randomization**.

Stratified randomization: Generic term for such procedures as **permuted blocks within strata** or the **minimization method**.

Stratified random sample: Sample which is obtained by **random sampling** or **probability sampling** for the **strata** in a **population**.

Stratified random sampling: First, a **stratification** is performed as described under the item stratification. Then, from each **stratum** a **random sample** is selected.

Stratified sample: See **stratification**.

Stratifying: Synonym for **stratification**.

Stratum: Subpopulation from a **population**, which contains all subjects of the population who exhibit the same level combination with respect to given **extraneous variables**. Also see **stratification**.

Strict random sample: From a defined population a **sample** is selected by means of a true random selection. In general, it is not possible to obtain strict random samples because in most cases the considered population is either too large or the selected subjects are not available for the researcher. In this case **limited random samples** have to be used alternatively.

Structural equality: This is given if different **samples** do not differ in their composition with respect to all known and unknown characteristics.

Study design: In most cases used as a synonym for **experimental design**, sometimes also used as generic term for experimental and **quasiexperimental design**.

Subjective endpoint: If subjective rating scales are used as **dependent variables**, these are called subjective endpoints.

Subject matching: This is present if in case of **matching** the experimental conditions are not randomly assigned to the subjects of a pair.

Subjects as their own control: If several experimental conditions are used in a **within-subjects design**, it is often claimed that here each subject serves as its own control. Here, the not very plausible assumption is made that no **carry-over effects** of experimental conditions occur, i.e. no effects of a condition which also affect measurements of the **dependent variable** after subsequent conditions. If such carry-over effects exist, a subsequent condition is applied to an altered subject which cannot be considered as a control. In particular, in most cases one cannot rule out that **asymmetric carry-over effects** occur. Here, the size of the post-effect of an experimental condition depends on the timely order in which the different conditions are applied. For the outcomes of such within-subjects designs, in general no **causal conclusions** are possible.

Sub-plot: See **plot**.

Suppressor effect: See **suppressor variable**.

Suppressor variable: An **independent variable** which is not related to the **dependent variable** but which is related to other independent variables. The **suppressor effect** works in that way that relations between the other independent variables and the dependent variable can be perceived only then if the influence of the suppressor variables is removed. Assume, e.g., that two different kinds of memory training are to be compared. As stimuli, pairs of words are used which consist always of one word in the mother tongue and another word with the same meaning in a foreign language. If no difference between the two training methods is detected, this might be caused by the fact that a part of the participants has some knowledge of the foreign language and, therefore, is able to assign the words in the correct way, irrespective of the particular training method. To eliminate the variable "knowledge of a foreign language", pairs of syllables without meaning should be used.

Surrogate endpoint: Sometimes a surrogate endpoint is used instead of the **endpoint** of interest. A surrogate endpoint is a **manifest variable** for which it is known that it is highly correlated with the endpoint of interest, i.e. for which it is known that a high linear relationship exists. A surrogate endpoint is used, if the endpoint of interest is difficult to measure or if its measurement is costly or if the measurement can be costly or if the measurement can be performed only at a far later point of time.

Surrogate observation: A **manifest variable** which is used as an **operationalization** of a **latent variable**.

Surrogate variable: If the clinical utility for a patient cannot be measured directly in a **clinical trial**, e.g., in case of quality of life, indirect criteria have to be used, so-called surrogate variables. I.e., a surrogate variable is a **manifest variable** which is derived from a **latent variable** by an **operationalization**.

Survey: See **census** or **sample survey**.

Switch back design: Synonym for **reversal design**.

Symmetrical block design: In a **block design**, t treatments with $r_1, ..., r_t$ **replications**, respectively, might occur. Further, b **blocks** with the sizes $s_1, ..., s_b$ might be present. Such a block design is called symmetric, if $t = b$ and if $(r_1, ..., r_t)$ might differ from $(s_1, ..., s_b)$ at most in the order of the elements. In the following example, we have $t = b = 3$, $r_1 = 3$, $r_2 = 2$, $r_3 = 1$ and $s_1 = 2$, $s_2 = 1$, $s_3 = 3$, where the 3 treatments are denoted by A, B, and C.

$$\boxed{\text{AB}}\quad\boxed{\text{A}}\quad\boxed{\text{ABC}}$$

Synchronic study: Study, where simultaneously occurring events are recorded.

Systematic allocation: In badly scheduled **clinical trials** sometimes a systematic allocation is used instead of a **random allocation**. E.g., patients might be allocated to the different treatment conditions according to the point of time they are accepted in the clinic, according to their date of birth, according to the initials of their name etc. For example, patients might be assigned to a new treatment if they were born in an even month (February, April etc.) and to a standard treatment if they were born in an odd month (January, March etc.).

Systematic assignment: Synonym for **systematic allocation**.

Systematic design: Design, where the experimental conditions are ordered in a regular scheme which is in most cases the same for all replications of the experiment. Because of the absent **randomization** no **causal conclusions** are possible.

Systematic error: An error of measurement which in most cases shows a bias in one specific direction. An example is given by a balance which always exhibits a weight that is too high.

Systematic observation: Synonym for **naturalistic observation study**.

Systematic sampling: A **population** with $N = kn$ subjects is considered which are consecutive-ly numbered from 1 to N. From this population a **sample** with n subjects is to be selected, where $1 < n < N$. A number m is randomly selected from the numbers 1, ..., k. The sample consists of the n subjects to whom the numbers m, $m + k$, $m + 2k$, ..., $m + (n - 1)k$ have been assigned. E.g., assume $k = 4$, $n = 3$ and, therefore, $N = 12$. From the numbers 1, 2, 3, and 4 the number 2 might have been randomly selected. Then, the subjects with the numbers 2, 6, and 10 belong to the sample.

Systematic variance: A part of the variance which is due to a cause which changes the measurements always in one direction. Assume, e.g., that the specific effect of a psychotherapy is to be established. In the **treatment group**, the patients are treated for a certain amount of time. In the **control group** the patients have a relaxed talk with a nurse for the same amount of time, where one assumes that the nurse has no training in psychotherapy. One has to assume that both treatments cause positive changes in the state of health of the patients. From this follows that a systematic variance is observed in the treatment group as well as in the control group, though the size of this variance might be different for both groups.

Systematizing the secondary variable: Synonym for **balancing**.

Tandem selection: See **selection**.

Target group: Synonym for **target population**.

Target population: That **population** for which the used **sample** is a **representative sample**. In general, this population is not known very well.

Target variable: Synonym for **primary variable**.

Tendency to agree: See **response bias**.

Tendency to the extremes: See **response bias**.

Testability: Causal conclusions about the behavior of subjects can only be drawn on the basis of observable **manifest variables.** This concerns **independent** as well as **dependent variables.** Causal conclusions about **latent variables** or **constructs** are not possible without an appropriate **operational definition.** Corresponding statements are only of a speculative nature and their truth cannot be tested. Here, the assumption of testability is not justified.

Testing: The effect of pretests in the sense of **reactivity, sensitization,** and **resistibility.**

Therapeutic trial: Synonym for **clinical trial.**

Third variable: A variable leading to an **illusory correlation** between two other variables. If the third variable which is also called **lurking variable** is kept fixed, the illusory correlation disappears.

Threat to validity: Each cause which yields **alternative explanations** for the outcomes of a study.

Three-period crossover design: A special case of the **extra period balanced crossover design** with the two **sequences** $A_1B_2B_3$ and $B_1A_2A_3$ for the two treatments A and B.

Three-point assay: Special three group design in drug studies used in order to compare a standard compound with a new compound. A first group receives a dose of the standard compound, a second group receives a dose of the new compound. This latter dose might have a size different from the former one. A third group gets no treatment or a **placebo.**

Threshold-crossing data: Measurement of that point of time, where a **dependent variable** crosses a given threshold.

Tied Latin squares: Assume that two **extraneous variables** are to be controlled with a **Latin square.** Assume further that the number of **levels** is equal to the number of experimental conditions for only one extraneous variable, while the number of levels is a multiple of the number of experimental conditions for the other extraneous variable. Then a design could be used which is generated by a random combination of the rows and columns of a corresponding number of Latin squares. In the following example we assume 3 treatments A, B, and C and 12 subjects. By a random arrangement of the columns of 4 Latin squares the design below is formed. The rows correspond to an extraneous variable with 3 levels. The original 4 Latin squares are given by the column combinations (8, 10, 3), (2, 9, 4), (11, 1, 7), and (6, 5, 12).

B	A	C	B	A	B	A	A	C	B	C	C
C	B	A	C	C	A	B	B	A	C	A	B
A	C	B	A	B	C	C	C	B	A	B	A

Time-dependent covariate: Covariate which is altered in the course of a study, e.g., the age of a subject during a long **prospective study.**

Time-independent covariate: Covariate which is not altered during a study in an essential or systematic way, e.g., the body size of a subject during a short study.

Time-lagged control design: Synonym for **multiple baseline design across subjects.**

Time-sampling studies: Longitudinal studies, where the measurements are not performed continuously or at equidistant points of time, but where these points of time are randomly selected.

Time-series design: In a one-group design first a **baseline** is recorded, then a treatment or **intervention** follows, and after this follow one or more posttests. Instead of only one intervention also several interventions might be scheduled which are separated from each other by sequences of measurements. It is permitted that different interventions differ only with respect to the point of time where they take place. Because these kinds of designs are **repeated-measures designs,** in general, no **causal conclusion** is possible.

Time-trade-off method: This is a technique for measuring preferences for different states of health, which was developed by Torrance et al. (1972) as an alternative to the **standard gamble.** Here, a patient is asked, whether he or she prefers to live for t years in an exactly defined state of disease, where t is the life expectancy for this state of disease, or whether he or she prefers to live only x years with $x < t$ but in the best of health. The time x is varied as long as the patient becomes indifferent with respect to the two alternatives for a certain time x_0. The preference value for the given state of disease is given by (x_0 / t).

Time-varying covariate: Synonym for **time-dependent covariate.**

Tolerance: Synonym for **tolerance level.**

Tolerance level: See **bioassay**.

Tolerance threshold: Synonym for **tolerance level**.

Tracking: The term tracking is used if the measurements of a subject are stable in time in a **within-subjects design**, i.e. subjects with high values at the beginning of a study exhibit high values during the whole study, while subjects with low values at the beginning of the study exhibit low values during the whole study.

Trait-treatment interaction: Interaction between a manipulated **independent variable** and a **cause variable**, e.g., gender or intelligence, which cannot be manipulated.

Transfer effect: Synonym for **carry-over effect**.

Treatment allocation ratio: The ratio of the two **sample sizes** if two treatments are compared with each other in a **clinical trial**.

Treatment condition: The experimental condition which is to be used in the **treatment group** or that condition in a **within-subjects design** for which a treatment is scheduled.

Treatment cross contamination: This is present if patients in a **clinical trial** do not receive the treatment which was assigned to them by a **random allocation** but one of the other treatments under study.

Treatment factor: A **factor** with **levels** which correspond to experimental conditions, i.e. a factor which is neither a **block factor** nor a **constant factor**.

Treatment group: Sample to which in contrast to a **control group** a treatment condition is assigned and where the effect of this condition on a **dependent variable** is of interest.

Treatment-period interaction: Synonym for **carry-over effect**.

Treatment-received analysis: In contrast to the **intention-to-treat analysis** and the **explanatory approach** patients in a **clinical trial** are considered for that group which corresponds to the actually received treatment and not for that group to which they were assigned by a **random allocation**. Here, **selection effects** cannot be ruled out.

Treatment-related attrition: Synonym for **differential attrition**.

Treatment-related refusals: A kind of **differential attrition** which occurs if subjects refuse to participate in a study after being informed to which treatment they were assigned.

Treatment sequence: Synonym for **sequence**.

Treatment trial: Synonym for **clinical trial**.

Treatment variable: Synonym for **independent variable**.

Trend: Systematic changes of measurements in the course of time which are observed for a **within-subjects design**. In particular, **monotonic increasing trends** are considered, where the measurement values increase in time, **monotonic decreasing trends**, where the measurement values decrease in time, and **cyclic trends**, where the measurement values first increase in time to reach a maximum, then decrease to a minimum, again increase in time to reach a maximum etc. (or vice versa).

Trial: Synonym for **experiment**.

Triangulation: Research strategy where more than one method is used to prove the existence of a causal relation. One way to achieve this aim is by means of different **operationalizations** of the **independent** and **dependent variables**.

Triple-blindfold experiment: See **blinding**.

Triple-blind study: See **blinding**.

Trivial block design: An **incomplete block design**, where the same number of experimental conditions is assigned to each **block** and where each possible combination of experimental conditions occurs exactly once as a block. For 4 experimental conditions where each block has to contain 2 experimental conditions the following trivial block design with 6 blocks results.

A	A	A	B	B	C
B	C	D	C	D	D

Trohoc: See **prolective cohort**.

Trojan square: Synonym for **Greco-Latin square**.

Truncated sample: In a censored sample measurements at certain subjects are not possible because these are not available before a certain point of time (**left-censored data**) or after a certain point of time (**right-censored data**) or during a time interval (**interval censored data**). A truncated sample is present if certain subpopulations are excluded when selecting subjects for a **sample**, i.e. for subjects from such subpopulations no measurements are available. If in a **clinical trial** only subjects with more than sixty years of age are considered, a **left-sided truncated sample** is present. If only subjects with less than 50 years of age are considered, a **right-sided truncated sample** is present. The terms censoring and truncation must not be mixed up with each other.

Two-armed bandit allocation: Special case of the **multi-armed bandit allocation** with only two treatment conditions.

Two-by-two crossover design: The design with the two **sequences** A_1B_2 and B_1A_2 which was described under the item **crossover design**.

Two-by-two design: Factorial design with two factors each with two **levels**.

Two-fold classification: See **double grouping**.

Two-fold cross-classification: See **cross-classification**.

Two-matched-groups design: Synonym for **match by correlated criterion design**.

Two-period design: Crossover design with two **periods**. Examples are the difficult to interpret **Latin square crossover design** with the sequences A_1B_2 and B_1A_2 and **Balaam's design** with the sequences A_1A_2, A_1B_2, B_1A_2, and B_1B_2.

Two-phase sampling: The selection of a **sample** is performed in two separated phases, where in the second phase, in contrast to **two-stage sampling**, another **dependent variable** is recorded. Here, in the first phase a large sample is selected on the basis of a first characteristic and then in a second phase a smaller subsample is drawn from this sample based on the outcomes of the first selection. In this subsample the second characteristic, in which the researcher is really interested, is recorded. This means that the first phase serves as a kind of **preselection**.

Two-point assay: Special two-groups design in drug studies to be able to estimate the effect of a compound. A first group receives a dose of the compound, a second group another dose.

Two-randomized-groups design: Synonym for **independent two group design**.

Two-stage sampling: The selection of a **sample** is performed in two separated phases. In contrast to **two-phase sampling** the same **dependent variable** is recorded in both phases. In the first phase the **population** is split up into large units, so-called **primary units** or **first-stage units**. From these a sample of units is selected. Each of the large units consists of smaller units, so-called **secondary units** or **second-stage units**. In the second phase from each of the large units a sample of the smaller units is selected. A special case is **cluster sampling**.

Two-stage stopping rule: In a **clinical trial** for each trial two sample sizes are fixed, a small one and a larger one. As soon as the smaller size is achieved for each treatment condition an interim analysis is performed. If the expected effects are then detected, the study is terminated. Otherwise, the study is continued until the larger sample size is achieved for each treatment condition. It should be observed that for this kind of procedure a special kind of sequential statistical evaluation is required.

Typical case sampling: A typical case sampling is a **purposive sampling** where only **sample units** are selected which are considered as particularly typical for the **population** in question, i.e. which cannot be considered as special cases.

Unaligned systematic sampling: A rectangular geographic region is subdivided into rectangles of the same size by drawing lines parallel to the edges. For each row consisting of rectangles an x-coordinate is randomly fixed which is the same for each rectangle of the row. Similarly, for each column consisting of rectangles a y-coordinate is randomly fixed which is the same for each rectangle of the column. For each rectangle a point within the rectangle is fixed by the given row and column coordinates. Now from each rectangle a subject is selected which is situated as near as possible to the fixed point.

Unbalanced design: Factorial design, for which not for all **level** combinations equal **sample sizes** are present.

Underadjustment: This is present if in case of a **statistical adjustment** a too small influence is ascribed to the **extraneous variables**.

Undersampling: See **oversampling**.

Unequal randomization: For statistical reasons one had better use **samples** of an as similar as possible size for the different treatment conditions. By this a higher efficiency of statistical procedures results which facilitates the detection of **effects**. If, however, e.g. the costs for different treatment conditions differ very much it might be advisable to select large samples for the treatment conditions with lower costs. This means for the **randomization** that the subjects are assigned to the different conditions with different probabilities.

Uniformity trial: Synonym for **dummy experiment**.

Unitary sampling: Subjects are being directly selected from the **population** and not, e.g., from a cluster as it is the case for **cluster sampling**.

Universe: Synonym for **population**.

Unobtrusive measure: Measuring device without **reactivity**.

Unobtrusive treatment: Subjects which are exposed to such a treatment do not know that the treatment is applied to them. Such treatments are favorable if, e.g., **hypothesis guessing** is to be avoided. For ethical reasons it is only rarely possible to realize unobtrusive treatments. An example might be that 30 out of 60 days of sale in a shop are randomly selected, where a certain kind of background music is played, while on the remaining 30 days another kind of music is used. The daily sales are used as a **dependent variable**.

Unrestricted random sampling: Synonym for **simple random sampling**.

Untreated control group design with pretest measures at more than one time interval: The **untreated control group design with pretest and posttest** is extended in that way that in both groups more than one **pretest** is scheduled (Cook and Campbell, 1979, pp. 117-118). Due to the absent randomization no causal conclusions are possible.

Untreated control group design with pretest and posttest: In two groups which were not generated by a **random assignment**, a **dependent variable** is measured twice. Here, in one group a treatment takes place between the two measurements (Cook and Campbell, 1979, pp. 103-112). Because of the absent **randomization** no **causal conclusions** are possible. Also compare the **before-after static group comparison design**.

Untreated control group design with proxy pretest measures: In two groups which were not generated by a **random assignment**, first one and then a second **dependent variable** is measured. In one of the two groups a treatment is placed between the two measurements. The dependent variable which is measured as the second one is the variable of interest. The dependent variable which is measured first should be as similar as possible to the second one (Cook and Campbell, 1979, pp. 112-115). Such a proxy pretest measure is used instead of the dependent variable of interest if either the dependent variable cannot be measured at the first point of time in a sound way or because of the **reactivity** of the dependent variable. E.g., in most cases it is not advisable to ask patients before a medical treatment whether they are content with the treating doctor. In many cases they will not even know the doctor at that point of time. However, if the patients know the doctor, such a question might have a reactive effect on the second question. A proxy pretest measure might consist in this case in a question concerning the patient's former experiences with doctors. **Causal conclusions** are not possible, e.g. because of the absent **randomization**.

Untreated control group design with separate pretest and posttest samples: In two groups which were not generated by a **random assignment**, each group is split up into two subsamples. First, the **dependent variable** is recorded at one subsample in each group. Then, in one of the two groups a treatment is applied. After this the dependent variable is recorded for those two subsamples in the two groups for which up to now no measurement was recorded (Cook and Campbell, 1979, pp. 115-117). In this case it is achieved that the existing **reactivity** of the measurement has no effect on the second measurement because this is recorded for other subjects. Because of the absent **randomization** no **causal conclusions** are possible.

Valid cases: Synonym for **per protocol population**.

Validity: Degree of exactness with which a measuring device informs about the **latent variable** to be measured.

Volunteer bias: If subjects are allowed to choose the experimental conditions which are applied to them themselves, i.e. if the conditions are not randomly applied to them, one cannot rule out that such subjects are highly motivated to exhibit a good performance or recovery, respectively.

Volunteer study: Study with healthy volunteers, e.g., in the first tests of a new drug in a **phase I study**.

Von Neumann-Morgenstern standard gamble: The **standard gamble** was first proposed by Von Neumann and Morgenstern (1953, pp. 17-19) to measure utility. In the context of the measurement of preferences for different states of health a patient might choose between two alternatives. According to the first alternative the patient will still live for exactly t years in a certain state of disease. According to the second alternative the patient will either achieve complete health and live for exactly t years and this with probability p or the patient will die immediately with probability $(1 - p)$. The probability p is varied as long as the patient becomes indifferent to the two alternatives. This might happen if a certain probability p_0 is used. The preference value for a given state of disease is given by p_0. Also compare the **time trade-off method**.

Waiting control group: It is often not possible, for ethical or practical reasons, to refuse a treatment to the subjects in a **control group**, e.g., to patients. Then, sometimes measurements are performed at patients who still have to wait for the treatment and these measurements are compared with the measurements at the treated subjects.

Wash-out period: In drug studies this term is used for the **rest period**.

Wave: A **panel**, i.e. a group of subjects, is interviewed several times in a **panel study**. These interviews are performed each time during a short time interval and are called waves. Longer time intervals might occur between the waves.

Whole-plot: See **plot**.

Withdrawal design: Synonym for **ABA design**.

Within-groups design: Synonym for **within-subjects design**.

Within-subjects design: In contrast to **between-subjects designs**, in within-subjects designs several experimental conditions are applied in a certain chronological order to the same subjects. In general, no **causal conclusions** are possible for the outcomes of within-subjects designs, because the effects of the different experimental conditions in most cases cannot be isolated.

Within-subjects factor: Synonym for **within-subjects variable**.

Within-subjects variable: An **independent variable** for which at each subject for at least two **levels** a **dependent variable** is recorded.

Yeasaying: A response tendency of subjects consisting in answering rather with "yes" than with "no", independent of the specific context. Also see **naysaying**.

Yoked control group design: If a potential **causal variable** is influenced by the behavior of the studied subjects, this renders **causal conclusions** more difficult. If certain kinds of behavior have certain consequences for the subjects in a **treatment group** which are scheduled in the **experimental design**, it might not be clear whether an increased behavior frequency is a reinforcement effect of the consequences as it is hoped or whether it is, e.g., simply an activation effect caused by the consequences. A yoked control group would be generated by forming pairs of subjects, where within each pair one subject is randomly assigned to the **treatment group** (**randomization**), however, the other one to the **control group**. If subjects of the treatment group exhibit the target behavior they experience the scheduled consequence. The subjects of the control group experience this consequence, independent of their own behavior, always, if the corresponding partner experiences the consequence. Thus both subjects of a pair experience the same number of consequences at the same points of time. If the frequency of the target behavior is higher for the subjects in the treatment group than the corresponding frequency for the subjects in the control group, one concludes that this increased frequency is caused by a reinforcement effect of the consequences. Church (1964) showed that this conclusion is not justified, i.e. that yoked control group designs cannot be considered as a solution of the problem that activity and reinforcement effects might be confounded.

Youden design: If in a **Latin square** only columns are deleted and if the resulting rows can be interpreted as the **blocks** of a **balanced incomplete block design**, a Youden design is formed (Youden, 1937, 1940). An example with 4 treatments A, B, C, and D is given in the following. By adding the fourth column (D, C, A, B) a Latin square results. In the 4 row blocks each pair of treatments occurs exactly twice. E.g., the treatment pair AD is found in the second and in the fourth row block.

A	B	C
B	A	D
C	D	B
D	C	A

Youden square: Synonym for **Youden design** though this is, by definition, not a square.

Zelen's single consent design: A proposal made by Zelen (1979) in order to address the problem that the requirement of **informed consent** often prevents a **random allocation** of patients in a **clinical trial**. According to this proposal a **sample** is selected from a **population**, where the given **inclusion** and **exclusion criteria** hold for the selected subjects. This sample is randomly split up (**randomization**) into a **control group** and a **treatment group**. The control group gets the usual standard treatment, such that informed consent is not necessary for these patients. The patients of the treatment group are asked whether they agree to a treatment with the new method (informed consent). Those patients who consent get the new treatment, the other patients of the treatment group get the standard treatment. The evaluation is performed by means of an **intention-to-treat analysis**, i.e. according to the random allocation and not according to the actually applied treatment.

Zero-order correlation: Correlation, i.e. a linear relationship between two variables, where no other variable has been kept constant. This might be an **illusory correlation**.

References

ADAMS, M. M. (1987). *William Ockham.* University of Notre Dame Press, Notre Dame, Indiana.

ATHENAEUS (1971). *The Deipnosophists. Vol I.* William Heinemann LTD, London.

BALAAM, L. N. (1968). *A two-period design with t^2 experimental units.* Biometrics 24, 61-73.

BATHER, J. (1980). *Randomised allocation of treatments in sequential trials.* Advances in Applied Probability 12, 174-182.

BERENBLUT, I. I. (1964). *Change-over designs with complete balance for first residual effects.* Biometrics 20, 707-712.

BERKSON, J. (1946). *Limitations of the application of fourfold table analysis to hospital data.* Biometrics Bulletin 2, 47-53.

BERNTSON, G. G., UCHINO, B. N., CACIOPPO, J. T. (1994). *Origins of baseline variance and the Law of Initial Values.* Psychophysiology 31, 204-210.

BIEFANG, S., KÖPCKE, W., SCHREIBER, M. A. (1979). *Manual für die Planung und Durchführung von Therapiestudien.* Springer, Berlin.

BOK, S. (1974). *The ethics of giving placebos.* Scientific American 231, 17-23.

BREWIN, T. B. (1982). *Consent to randomised treatment.* Lancet, 919-921.

BROWN, B. W. (1980). *The crossover experiment for clinical trials.* Biometrics 36, 69-79.

BRUNNER, E., DENKER, M. (1994). *Rank statistics under dependent observations and applications to factorial designs.* Journal of Statistical Planning and Inference 42, 353-378.

BRUNNER, E., PURI, M. L., SUN, S. (1995). *Nonparametric methods for stratified two-sample designs with application to multiclinic trials.* Journal of the American Statistical Association 90, 1004-1014.

BURKHARDT, R., KIENLE, G. (1978). *Controlled clinical trials and medical ethics.* Lancet, 1356-1359.

CAMPBELL, D. T. (1957). *Factors relevant to the validity of experiments in social settings.* Psychological Bulletin 54, 297-312.

CARRIÈRE, K. C. (1994). *Crossover designs for clinical trials.* Statistics in Medicine 13, 1063-1069.

CARRIÈRE, K. C., REINSEL, G. C. (1992). *Investigation of dual-balanced crossover designs for two treatments.* Biometrics 48, 1157-1164.

CHURCH, R. M. (1964). *Systematic effect of random error in the yoked control design.* Psychological Bulletin 62, 122-131.

COOK, T. D., CAMPBELL, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Rand McNally College Publishing Company, Chicago.

CROMBIE, A. C. (1952). *Avicenna's influence on the medieval scientific tradition.* In: WICKENS, G. M. (ed.), *Avicenna: Scientist and Philosopher, A Millenary Symposium.* pp. 84-107. Luzac and Co., London.

DILSAVER, S. C., MAJCHRZAK, M. J. (1990). *Effects of placebo (saline) injections on core temperature in the rat.* Progress in Neuro-Psychopharmacology & Biological Psychiatry 14, 417-422.

DORFMAN, R. (1943). *The detection of defective members of large populations.* Annals of Mathematical Statistics 14, 436-440.

DOWNING, R. W., RICKELS, K. (1980). *Responders and nonresponders to chlordiazepoxide and placebo: A discriminant function analysis.* Progress in Neuro-Psychopharmacology 4, 405-415.

EBBINGHAUS, H. (1885). *Über das Gedächtnis: Untersuchungen zur Experimentellen Psychologie.* Duncker & Humblot, Leipzig.

EDGINGTON, E. S. (1967). *Statistical inference from N = 1 experiments.* The Journal of Psychology 65, 195-199.

EDGINGTON, E. S. (1972a). *An additive method for combining probability values from independent experiments.* The Journal of Psychology 80, 351-363.

EDGINGTON, E. S. (1972b). *N = 1 experiments: hypothesis testing.* The Canadian Psychologist 13, 121-134.

EDGINGTON, E. S. (1975). *Randomization tests for one-subject operant experiments.* The Journal of Psychology 90, 57-68.

EDGINGTON, E. S. (1980a). *Random assignment and statistical tests for one-subject experiments.*Behavioral Assessment 2, 19-28.

EDGINGTON, E. S. (1980b). *Validity of randomization tests for one-subject experiments.* Journal of Educational Statistics 5, 235-251.

EDGINGTON, E. S. (1980c). *Overcoming obstacles to single-subject experimentation.* Journal of Educational Statistics 5, 261-267.

EDGINGTON, E. S. (1982). *Nonparametric tests for single-subject multiple schedule experiments.* Behavioral Assessment 4, 83-91.

EDGINGTON, E. S. (1984). *Statistics and single case analysis.* Progress in Behavior Modification 16, 83-119.

EDGINGTON, E. S. (1987). *Randomized single-subject experiments and statistical tests.* Journal of Counseling Psychology 34, 437-442.

EDGINGTON, E. S. (1992). *Nonparametric tests for single-case experiments.* In: KRATOCHWILL, T. R., LEVIN, J. R. (eds.): *Single-case Research Design and Analysis: New Directions for Psychology and Education.* pp. 133-157. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

EDGINGTON, E. S. (1995). *Randomization Tests. Third Edition.* Marcel Dekker, Inc., New York.

EDGINGTON, E. S. (1996). *Randomized single-subject experimental designs.* Behaviour Research and Therapy 34, 567-574.

EDGINGTON, E. S., BLAND, B. H. (1993). *Randomization tests: application to single-cell and other single-unit neuroscience experiments.* Journal of Neuroscience Methods 47, 169-177.

EFRON, B. (1971). *Forcing a sequential experiment to be balanced.* Biometrika 58, 403-417.

ERISMANN, T. (1921). *Psychologie. Volume II: Die allgemeinsten Eigenschaften der Psyche.* Walter de Gruyter & Co., Berlin and Leipzig.

EVERITT, B. S. (1995). *The Cambridge Dictionary of Statistics in the Medical Sciences.* Cambridge University Press, New York.

FECHNER, G. T. (1860). *Elemente der Psychophysik. First Part.* Breitkopf & Härtel, Leipzig.

FEINSTEIN, A. P. (1973). *Clinical biostatistics: XX. The epidemiologic trohoc, the ablative risk ratio, and 'retrospective' research.* Clinical Pharmacology and Therapeutics 14, 291-307.

FERSTER, C. B., SKINNER, B. F. (1957). *Schedules of Reinforcement.* Appleton-Century-Crofts, Inc., New York.

FILE, S. E. (1992). *Effects of lorazepam on psychomotor performance: A comparison of independent-groups and repeated-measures designs.* Pharmacology Biochemistry and Behavior 42, 761-764.

FISHER, R. A. (1925). *Statistical Methods for Research Workers. First Edition.* Oliver and Boyd, Edinburgh.

FISHER, R. A. (1926). *The arrangement of field experiments.* Journal of the Ministry of Agriculture, Fisheries and Food 33, 503-513.

FISHER, R. A. (1966). *The Design of Experiments. Eighth Edition.* (First Edition 1935). Oliver and Boyd, Edinburgh.

FLEISS, J. L. (1989). *A critique of recent research on the two-treatment crossover design.* Controlled Clinical Trials 10, 237-243.

FREEDMAN, S. R., ENRIGHT, R. D. (1996). *Forgiveness as an intervention goal with incest survivors.* Journal of Consulting and Clinical Psychology 64, 983-992.

FREEMAN, P. R. (1989). *The performance of the two-stage analysis of two-treatment, two-period crossover trials.* Statistics in Medicine 8, 1421-1432.

FRIEDMAN, B. (1949). *A simple urn model.* Communications on Pure and Applied Mathematics 2, 59-70.

GALTON, F. (1877). *Typical laws of heredity.* Proceedings of the Royal Institution of Great Britain 8, 281-301.

GALTON, F. (1885). *Regression towards mediocrity in hereditary stature.* Journal of the Anthropological Institute of Great Britain and Ireland 15, 246-263.

GILBERT, J. P., McPEEK, B., MOSTELLER, F. (1977). *Statistics and ethics in surgery and anesthesia.* Science 198, 684-689.

GØTZSCHE, P. C. (1994). *Is there logic in the placebo?* Lancet 344, 925-926.

GREENWALD, A. G. (1976). *Within-subjects designs: To use or not to use?* Psychological Bulletin 83, 314-320.

GRICE, G. R. (1966). *Dependence of empirical laws upon the source of experimental variation.* Psychological Bulletin 66, 488-498.

GRICE, G. R., HUNTER J. J. (1964). *Stimulus intensity effects depend upon the type of experimental design.* Psychological Review 71, 247-256.

GRIZZLE, J. E. (1965). *The two-period change-over design and its use in clinical trials.* Biometrics 21, 467-480. Corrigenda in Biometrics 30 (1974) 727.

GRÜNBAUM, A. (1986). *The placebo concept in medicine and psychiatry.* Psychological Medicine 16, 19-38.

GUILFORD, J. P. (1954). *Psychometric Methods. Second Edition.* McGraw Hill, New York.

HARDEN, R. N., GRACELY, R. H., CARTER, T., WARNER, G. (1996). *The placebo effect in acute headache management: Ketorolac, meperidine, and saline in the emergency department.* Headache 36, 352-256.

HERODOTUS (1866). *Die Musen. Volume 2: Euterpe.* Translated by J. Chr. F. Bähr. Krais & Hoffmann, Stuttgart.

HILL, A. B. (1963). *Medical ethics and controlled trials.* British Medical Journal, 1043-1049.

HOLLAND, P. W. (1986). *Statistics and causal inference.* Journal of the American Statistical Association 81, 945-960.

HOLLINGWORTH, H. L. (1912). *The influence of caffein alkaloid on the quality and amount of sleep.* American Journal of Psychology 23, 89-100.

HOVLAND, C. I., SHERIF, M. (1952). *Judgmental phenomena and scales of attitude measurement: Item displacement in Thurstone scales.* Journal of Abnormal and Social Psychology 47, 822-832.

ISAAC, W. L., ISAAC, W. (1977). *Differences in placebo effects.* Pharmacology Biochemistry and Behavior 6, 235-236.

KANT, I. (1781). *Critik der reinen Vernunft. First Edition.* Johann Friedrich Hartknoch, Riga.

KLING, J. W., HOROWITZ, L., DELHAGEN, J. E. (1956). *Light as a positive reinforcer for rat responding.* Psychological Reports 2, 337-340.

KNAPP, T. R. (1982). *A case against the single-sample repeated-measures experiment.* Educational Psychologist 17, 61-65.

LASKA, E. M., MEISNER, M. (1985). *A variational approach to optimal two-treatment crossover designs: Application to carryover-effect models.* Journal of the American Statistical Association 80, 704-710.

LASKA, E., MEISNER, M., KUSHNER, H. B. (1983). *Optimal crossover designs in the presence of carryover effects.* Biometrics 39, 1087-1091.

LENTZ, T. F. (1938). *Acquiescence as a factor in the measurement of personality.* Psychological Bulletin 35, 659.

LIPPS, G. F. (1921). *Grundriss der Psychophysik. Third Edition.* Walter de Gruyter & Co., Berlin und Leipzig.

LUTHER, M. (1862). *Die Bibel oder die ganze Heilige Schrift des alten und neuen Testaments, nach der deutschen Übersetzung Dr. Martin Luthers.* Wilhelm Hassel, Cöln.

MACKINTOSH, N. J. (1977). *Stimulus control: Attentional factors.* In: Honig, W. K., Staddon, J. E. R. (eds.), *Handbook of Operant Behavior.* pp. 481-513. Prentice-Hall, Englewood Cliffs, New Jersey.

MATHESON, D. W., BRUCE, R. L., BEAUCHAMP, K. L. (1971). *Introduction to Experimental Psychology.* Holt, Rinehart and Winston, London.

MATHEWS, C. O. (1927). *The effect of position of printed response words upon children's answers to questions in two-response types of tests.* Journal of Educational Psychology 18, 445-457.

MATTHEWS, J. N. S. (1990). *Optimal dual-balanced two treatment crossover designs.* Sankhyā: The Indian Journal of Statistics 52, Series B, 332-337.

McGUIGAN, F. J. (1978). *Experimental Psychology: a Methodological Approach. Third Edition* (First Edition 1960). Prentice-Hall, Englewood Cliffs, New Jersey.

McINTYRE, G. A. (1952). *A method of unbiased selective sampling, using ranked sets.* Australian Journal of Agricultural Research 3, 385-390.

MELLERS, B. A., DAVIS, D. M., BIRNBAUM, M. H. (1984). *Weight of evidence supports one operation for "ratios" and "differences" of heaviness.* Journal of Experimental Psychology: Human Perception and Performance 10, 216-230.

MIKÉ, V. (1989). *Philosophers assess randomized clinical trials: The need for dialogue.* Controlled Clinical Trials 10, 244-253.

MILL, J. S. (1846). *System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation. Second Edition.* John W. Parker, West Strand, London.

PAVLIK, W. B., CARLTON, P. L. (1965). *A reversed partial-reinforcement effect.* Journal of Experimental Psychology 70, 417-423.

PAVLOV, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex.* Oxford University Press, Oxford.

PEIRCE, C. S., JASTROW, J. (1885). *On small differences of sensation.* Memoirs of the National Academy of Sciences for 1884, 3, 75-83.

PFUNGST, O. (1977). *Der kluge Hans, ein Beitrag zur nicht-verbalen Kommunikation.* (Reprint of: Das Pferd des Herrn von Osten (Der kluge Hans): Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie, Barth, Leipzig, 1907.) Fachbuchhandlung für Psychologie, Frankfurt am Main.

POCOCK, S. J. (1979). *Allocation of patients to treatment in clinical trials.* Biometrics 35, 183-197.

POCOCK, S. J., SIMON, R. (1975). *Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial.* Biometrics 31, 103-115.

POLITZ, A., SIMMONS, W. (1949). *An attempt to get the "not at homes" into the sample without callbacks.* Journal of the American Statistical Association 44, 9-31.

POLITZ, A., SIMMONS, W. (1950). *Note on "An attempt to get the not-at-homes into the sample without callbacks".* Journal of the American Statistical Association 45, 136-137.

POULTON, E. C. (1973). *Unwanted range effects from using within-subject experimental designs.* Psychological Bulletin 80, 113-121.

POULTON, E. C. (1974). *Range effects are characteristic of a subject serving in a within-subjects experimental design – A reply to Rothstein.* Psychological Bulletin 81, 201-202.

RASCH, D., TIKU, M. L., SUMPF, D. (1994). *Elsevier's Dictionary of Biometry.* Elsevier, Amsterdam.

REITER, J. (2000). *Using statistics to determine causal relationships.* American Mathematical Monthly 107, 24-32.

RIECKEN, H. W., BORUCH, R. F., CAMPBELL, D. T., CAPLAN, W., GLENNAN, T. K., PRATT, J. W., REES, A., WILLIAMS, W. (1974). *Social Experimentation: A Method for Planning and Evaluating Social Intervention.* Academic Press, New York.

ROBBINS, H. (1952). *Some aspects of the sequential design of experiments.* Bulletin of the American Mathematical Society 58, 527-535.

ROBINSON, P. W. (1976). *Fundamentals of Experimental Psychology: A Comparative Approach.* Prentice-Hall, Englewood Cliffs, New Jersey.

ROSENTHAL, R. (1966). *Experimenter Effects in Behavioral research.* Appleton-Century-Crofts, New York.

ROSENTHAL, R., FODE, K. L. (1963). *The effect of experimenter bias on the performance of the albino rat.* Behavioral Science 8, 183-189.

ROSENTHAL, R., LAWSON, R. (1964). *A longitudinal study of the effects of experimenter bias on the operant learning of laboratory rats.* Journal of Psychiatric Research 2, 61-72.

ROSENZWEIG, P., BROHIER, S., ZIPFEL, A. (1993). *The placebo effect in healthy volunteers: Influence of experimental conditions on the adverse events profile during phase I studies.* Clinical Pharmacology and Therapeutics 54, 578-583.

ROYALL, R. M. (1991). *Ethics and statistics in randomized clinical trials.* Statistical Science 6, 52-88.

RUBIN, D. B. (1974). *Estimating causal effects of treatments in randomized and nonrandomized studies.* Journal of Educational Psychology 66, 688-701.

RUBIN, D. B. (1978). *Bayesian inference for causal effects: the role of randomization.* Annals of Statistics 6, 34-58.

RUBIN, D. B. (1990). *Formal modes of statistical inference for causal effects.* Journal of Statistical Planning and Inference 25, 279-292.

SELIGMAN, M. E. P., BEAGLEY, G. (1975). *Learned helplessness in the rat.* Journal of Comparative and Physiological Psychology 88, 534-541.

SENN, S. (1994). *Fisher's game with the devil.* Statistics in Medicine 13, 217-230.

SHAH, M. H. (1966). *The General Principles of Avicenna's Canon of Medicine.* Naveed Clinic, Karachi.

SHAPIRO, A. K., MORRIS, L. A. (1978). *The placebo effect in medical and psychological therapies.* In: GARFIELD, S. L., BERGIN, A. E. (eds.), *Handbook of Psychotherapy and Behavior Change: An Empirical Analysis. Second Edition.* pp. 369-410. John Wiley and Sons, New York.

SKINNER, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis.* Appleton-Century-Crofts, New York.

SKINNER, B. F. (1961). *Cumulative Record. Enlarged Edition.* Appleton-Century-Crofts, New York.

SOLOMON, R. L. (1949). *An extension of control group design.* Psychological Bulletin 46, 137-150.

STRAUS, J. L., VON AMMON CAVANAUGH, S. (1996). *Placebo effects: Issues for clinical practice in psychiatry and medicine.* Psychosomatics 37, 315-326.

THOMPSON, W. R. (1933). *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.* Biometrika 25, 285-294.

THOMPSON, W. R. (1935). *On the theory of apportionment.* American Journal of Mathematics 57, 450-456.

THORNDIKE, E. L. (1920). *A constant error in psychological ratings.* Journal of Applied Psychology 4, 25-29.

TORRANCE, G. W., THOMAS, W. H., SACKETT, D. L. (1972). *A utility maximization model for evaluation of health care programs.* Health Services Research 7, 118-133.

UNDERWOOD, B. J. (1957). *Interference and forgetting.* Psychological Review 64, 49-60.

UNDERWOOD, B. J., SHAUGHNESSY, J. J. (1975). *Experimentation in Psychology.* John Wiley & Sons, New York.

VOGT, W. P. (1993). *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences.* SAGE Publications, Newbury Park, London, New Delhi.

VON NEUMANN, J., MORGENSTERN, O. (1953). *Theory of Games and Economic Behavior. Third Edition* (First Edition 1944). Princeton University Press, Princeton.

WEI, L. J., DURHAM, S. (1978). *The randomized play-the-winner rule in medical trials.* Journal of the American Statistical Association 73, 840-843.

WILDER, J. (1931). *Das „Ausgangswert-Gesetz" – ein unbeachtetes biologisches Gesetz; seine Bedeutung für Forschung und Praxis.* Klinische Wochenschrift 10. Jahrgang, Nr. 41, 1889-1893.

WOLF, S., PINSKY, R. H. (1954). *Effects of placebo administration and occurrence of toxic reactions.* Journal of the American Medical Association 155, 339-341.

WUNDT, W. (1911). *Grundriss der Psychologie. 10th Edition.* (First Edition 1896). Verlag von Wilhelm Engelmann, Leipzig.

WURTHMANN, C., KLIESER, E., LEHMANN, E., KRAUTH, J. (1996). *Single-subject experiments to determine individually differential effects of anxiolytics in generalized anxiety disorder.* Neuropsychobiology 33, 196-201.

YATES, F. (1938). *The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments.* Biometrika 30, 440-466.

YOUDEN, W. J. (1937). *Use of incomplete block replications in estimating tobacco-mosaic virus.* Contributions from the Boyce Thompson Institute 9, 41-48.

YOUDEN, W. J. (1940). *Experimental designs to increase accuracy of greenhouse studies.* Contributions from the Boyce Thompson Institute 11, 219-228.

ZELEN, M. (1969). *Play the winner rule and the controlled clinical trial.* Journal of the American Statistical Association 64, 131-146.

ZELEN, M. (1979). *A new design for randomized clinical trials.* The New England Journal of Medicine 300, 1242-1245.

ZIEGLER, H. E. (1921). *Tierpsychologie.* Walter de Gruyter & Co., Berlin.

ZIFFERBLATT, S. M., WILBUR, C. S. (1978). *Commentary: A psychological perspective for double-blind trials.* Clinical Pharmacology and Therapeutics 23, 1-10.

Author Index

Adams – 17
Athenaeus – 3, 13, 152
Balaam – 212, 223
Bather – 244
Beagley – 71
Beauchamp – viii, 79, 93
Berenblut – 214
Berkson – 214
Berntson – 27
Biefang – 39, 53
Birnbaum – 85
Bland – 143
Bok – 68
Brewin – 45
Brohier – 69
Brown – 98
Bruce – viii, 79, 93
Brunner – 24
Burkhardt – 45
Campbell – viii, 21f., 29, 218, 221, 227, 230, 235, 237, 246, 252, 258, 261, 272
Carrière – 98f.
Carlton – 84
Church – 72f., 273
Cook – viii, 21f., 29, 218, 221, 227, 230, 235, 237, 246, 252, 258, 261, 272
Crombie – 3
Davies – 85
Delhagen – 50, 71f.
Denker – 24
Dilsaver – 69
Dorfman – 225
Downing – 69
Durham – 47, 252
Ebbinghaus – 84
Edgington – 143ff., 149, 203, 226
Efron – 214, 250f.
Enright – 72
Erismann – 9f., 13
Everitt – viii
Fechner – 7ff.
Feinstein – 254
Ferster – 71
File – 85
Fisher – 7f., 10f., 13, 37, 141, 143ff., 149, 226, 239
Fleiss – 98
Fode – 57
Freedman – 72
Freeman – 98
Friedman – 232
Galton – 27
Gilbert – 45
Gøtzsche – 69
Greenwald – 85
Grice – 84f.
Grizzle – 98
Grünbaum – 69
Guilford – 260
Harden – 69
Herodotus – 81f.
Hill – 45
Holland – 16
Hollingworth – 69
Horowitz – 50, 71f.
Hovland – 260
Hunter – 84
Isaac – 69
Jastrow – 7ff., 13
Kant – 18
Kienle – 45
Kling – 50, 71f.
Knapp – 85
Köpcke – 39, 53
Kushner – 98
Laska – 98
Lawson – 57
Lentz – 260
Lipps – 9
Luther – 2
Mackintosh – 50, 241, 249
Majchrzak – 69
Matheson – viii, 79, 93, 213f., 220, 241, 248, 253, 258, 262, 266
Mathews – 260
Matthews – 99
McGuigan – viii, 54f., 220, 228f., 263
McIntyre – 257
McPeek – 45
Meisner – 98
Mellers – 84
Mill – 4ff., 13, 153f.
Miké – 45
Moekel – 31
Morgenstern – 273
Morris – 69
Mosteller – 45
Pavlik – 84
Pavlov – 50
Peirce – 7ff., 13
Pfungst – 31, 57, 167
Pinsky – 69
Pocock – 215, 242
Politz – 252
Poulton – 85
Puri – 24
Rasch – viii
Reinsel – 98
Reiter – 16
Rickels – 69
Riecken – 46, 170f.
Robbins – 244
Robinson – 30, 211, 229, 266
Rosenthal – 30f., 57, 73ff., 165f., 229, 255
Rosenzweig – 69
Royall – 45
Rubin – 16
Schreiber – 39, 53
Seligman – 71
Senn – 142
Shah – 3
Shapiro – 69
Shaughnessy – 55, 219, 235, 254
Sherif – 260
Simmons – 252
Simon – 242
Skinner – 57f., 71
Solomon – 77, 131f., 196
Straus – 69
Sumpf – viii
Sun – 24
Tiku – viii
Thompson – 244
Thorndike – 260

Torrance – 269
Underwood – 55, 84, 219, 235, 254
Vogt – viii
von Ammon Cavanaugh – 69
von Neumann – 273
Wei – 47, 252
Wilbur – 66
Wilder – 27, 240
Wolf – 69
Wundt – 6, 13, 43, 155
Wurthmann – 141
Yates – 11f.
Youden – 274
Zelen – 46f., 171, 244, 251f., 256, 274
Ziegler – 31, 57f.
Zifferblatt – 66
Zipfel – 69

Subject Index

Absolute effect – 188
Adaptive design – 47
Adverse effect – 15
Alpha-adjustment – 24
Alternative explanation – 21
Analysis of covariance – 86
Analysis of variance – 103, 118
Arbitrariness – 7, 155
Artifact – 16, 27
Assumptions of statistical tests – 23
Asymmetric transfer – 61
Balancing – 58f., 64
Baseline – 82, 125
Before-after design – 125
Bioequivalence problem – 23
Blinding – 65, 142
Block – 11, 51
Blocking – 50f., 173
Blocking off – 48
Block variable – 11, 50
Bonferroni-correction – 24
Camera silens – 48
Carry-over effect – 60, 96, 98f.
Causal chain – 17, 159, 161
Causal conclusion – 28
Causal relation – 15f., 19, 159
Causal variable – 15
Ceiling effect – 26, 33
Change-over design – 99
Clever Hans – 31, 166f.
COD (t, p, s) – 99
Code – 66
Comparability – 39
Comparison group – 78
Compensatory equalization of treatments – 29
Compensatory rivalry – 29
Complete balancing – 60
Complete counterbalancing – 61
Complete expectancy control – 74
Confounding – 21
Conservative arrangement of the levels of the extraneous variable – 80
Constancy – 49
Construct – 17f., 29
Construct validity – 22
Control – 37
Control condition – 2, 67, 151
Control group – 2f., 67, 70, 93, 109
Correlative relation – 19, 159
Counterbalancing – 60, 64
Covariate – 86
Covariation – 16
Covering – 49f.
Crossover design – 64, 95, 99
Dark room effect – 165
Dependent variable – 15
Difference score – 76, 85
Differential effect – 102, 112
Differential transfer – 61
Diffusion of treatments – 29
Dose-response curve – 15, 112
Double-blind study – 30, 66, 166
Dropout – 128
Dropout rate – 28, 195
Edgington experiment – 203
Effect variable – 15
Elimination – 48
Error variance – 25, 99
Ethics – 45
Expectancy control group – 73
Experiment – 79
Experimental group – 2
Experimental mortality – 28, 128, 137
Experimenter expectancies – 30
External validity – 22, 161
Extraneous variable – 16, 25, 37, 48, 55
Extreme groups – 53
Factor – 15
Factorial design – 11, 40, 103
Factor level – 15
First-order interaction – 117
Fixed factor – 15
Floor effect – 26, 34
Gain score – 85, 94
Generalizability – 22, 33
Global control – 50
Greco-Latin square – 63
Habituation – 26
Hawthorne effect – 30
Historical control – 79
History – 25, 28, 34
Hypothesis guessing – 30
Illusory correlation – 19
Imitation of treatments – 29
Incomplete balancing – 60
Incomplete counterbalancing – 61
Independent variable – 15,55
Indirect Method of Difference – 5, 153
Instrumentation – 26, 28
Intention-to-treat – 46
Interaction – 33f., 54, 75, 77, 101f., 112, 184
Internal validity – 22, 39, 161
Interrupted time-series design – 145
Intervening variable – 17
Intervention – 82, 125
Intervention study – 203
Irreversible effect – 61
Isolation – 7
John Henry Effect – 29
Joint Method of Agreement and Difference – 5
Latent variable – 19
Latin square – 11, 62
Law of Initial Values – 27
Level – 15, 158
Local control – 50
Local randomization – 51
Main effect – 100f., 117
Manifest variable – 19
Manipulation – 7
Masking – 50
Matching – 50f., 173
Matching variable – 50
Maturation – 26, 28
Measurement error – 16, 27
Method of Agreement – 5, 153
Method of Concomitant Variations – 6, 154
Method of Difference – 5, 153
Method of Residues – 5, 154
Mono-method bias – 30
Mono-operation bias – 29
Multifactorial design – 101
Multiple tests – 24
Multiple treatments – 40
Nocebo – 69
Nonequivalent comparison group – 78
Nonequivalent control group – 78
Null result – 23, 161
Observational equality – 39
Occam's razor – 17
Ockham's razor – 17, 160
One-factor design – 109
One-group posttest only design – 124

One-group pretest-posttest design – 124
Open protocol – 142
Operational equality – 40
Operationalization – 17
Optimal design – 99
Overadjustment – 85
Over-matching – 53
Overshadowing – 50
Pairing – 50
Partial expectancy control – 74
Percentage change – 76, 85
Period – 99
Pilot experiment – 89
Pilot study – 89, 179
Placebo – 32, 68, 70, 110
Play-the-winner rule – 47
Population – 37
Posttest – 32, 76
Power – 22, 25, 161
Precision – 24, 95
Precision of a design – 182
Preliminary experiment – 32, 89, 179
Pre-post design – 125
Pretest – 32, 54, 76
Principle of Parsimony – 17f.
Principle of Testability – 18
Pseudorandom number – 38
Qualitative factor – 25
Quantitative factor – 25
Quasiexperimental design – 78
Random difference – 25
Random disturbance – 25
Random factor – 15
Randomization – 2, 7, 10f., 23, 37, 39f., 45, 62, 92, 151, 156f.
Randomization test – 143
Randomized blocks – 52
Randomized pairs – 50
Randomized play-the-winner rule – 47
Random sample – 37
Random number – 11
Regression artifact – 27
Regression to the mean – 27
Regression towards mediocrity – 27
Relative effect – 188
Reliability – 24f.
Repeated measures – 82ff., 94, 124
Repeated measures analysis of variance – 84
Repeated-measures design – 23, 95, 125, 194f., 197
Repeated-measures factor – 99
Replication – 3, 56
Representative equality – 40
Representativeness – 39
Reproducibility – 7
Resentful demoralization – 29
Residual effect – 98
Residues – 85
Response variable – 15
Reversion – 27
Rosenthal effect – 30, 165
Sampling with replacement – 37
Sampling without replacement – 37
Second-order interaction – 117
Selection – 28, 33
Selection effect – 28, 83, 195
Selection of the levels of the independent variable – 16
Sensitization – 26
Sequence – 99
Setting – 34
Sham lesion – 110
Sham operation – 110
Significance test – 23
Single-blind study – 30, 66, 166
Single-case experiment – 141, 143
Single-case study – 203
Social desirability responding – 30, 164
Solomon four group design – 77, 132, 196
Solomon three group design – 78
Standardization of treatment – 25
Statistical adjustment – 85
Statistical conclusion validity – 22
Statistical regression – 27
Statistical test – 23f., 161
Statistical twins – 51
Strata – 52
Structural equality – 39
Test – 23
Testing – 26
Third variable – 19, 28
Threats to validity – 21
Time-series design – 82, 125, 145
Transfer effect – 60, 96, 98
Treatment group – 67, 109
Trend – 128
Triple-blind study – 31, 66, 166
Underadjustment – 85
Waiting control group – 109
Wash-out period – 83, 98, 127
Within-groups design – 83
Yoked box – 71
Yoked control groups – 71