

Chapter 2

Experimental Design

Abstract This chapter starts with explaining the difference between an experiment and a quasi-experiment. Next, between-subjects and within-subject research designs are compared, and criteria about the choice for either design are discussed. The importance of a control group is highlighted, and techniques for participant assignment to groups are presented. Validity threats are described, including sample representativeness, demand characteristics, experimenter expectancy bias, causation versus correlation, and attrition. We explain the notion of statistical reliability and discuss self-reported measures and associated pitfalls such as social desirability and response style.

2.1 Experiments and Quasi-experiments

Before starting to collect data, it is important to decide whether to conduct an *experiment* or a *quasi-experiment*. An experiment is a study in which a researcher exposes one or more participant groups to an intervention and investigates the effects of this intervention. That is, in an experiment, the exposure of participant groups to the intervention is controlled by the researcher. Experiments can be between-subjects, where two or more independent participant groups undergo different procedures or perform different tasks (i.e., are tested in different conditions), or within-subject, where each participant undergoes multiple procedures or performs multiple tasks. A mixed factorial design is also possible. In a mixed design, two or more independent variables (i.e., the variables that are systematically manipulated by the researcher; in other words, the intervention) are investigated, at least one of which is manipulated between-subjects and at least another one is manipulated within-subject. When it is practically, financially, or ethically undesirable to conduct an experiment, a quasi-experiment can be done. In a quasi-experiment, the exposure of participant groups to the intervention is not controlled by the researcher. There are several types of quasi-experiments; in this chapter, cohort studies and case-control studies will be discussed.

2.2 Between-Subjects Design

In a between-subjects design, each participant is assigned to a group. The groups are treated in the exact same way, except that each group undergoes a different procedure or performs a different task.

Between-subjects experiments are common in medical research for testing the efficacy of a treatment as compared to not receiving the treatment (Bhatt 2010; Connors et al. 1996). Between-subjects experiments are also used in human subject research in engineering. For example, a researcher may conduct a between-subjects experiment to compare the effects of feedback on task performance (e.g., in Zanotto et al. 2013, participants were assigned to one of four groups; each group received a different type of feedback while walking with an exoskeleton, and the gait characteristics of the four groups were compared).

2.2.1 Control Groups

In a randomized controlled trial, at least one of the groups is a control group. There are several types of control groups: (1) a control group in which participants undergo a sham procedure or perform a sham task (also called *placebo control group* or *negative control group*), (2) a control group in which participants undergo a procedure or perform a task that has been previously tested and has a known effect on the outcome variable (*positive control group*), or (3) a control group in which participants do nothing (*natural history control group*).

Randomized controlled trials are the gold standard for making causal inferences about the effect of the experimental treatment on one or more outcome measures (Guyatt et al. 2008; National Health and Medical Research Council 1999; U.S. Preventive Services Task Force 1996). It is also possible to conduct experiments without a control group; such experiments are called randomized trials rather than randomized *controlled* trials.

Textbox 2.1 The placebo effect and its use in human subject research

A placebo is a sham procedure/task/device that simulates a real procedure/task/device. While a placebo is supposed to be ineffective, it is often that humans respond to it (De Craen et al. 1999). This phenomenon is called the *placebo effect* and may relate to expectancies about the effect of the real procedure/task/device. Interestingly, the placebo effect is not merely a subjective impression but can also lead to actual physiological changes (Oken 2008). For example, it has been shown that placebo caffeine induces dopaminergic responses (measured with positron emission tomography; Kaasinen et al. 2004) and physiological arousal (Mikalsen et al. 2001) comparable to the corresponding effects induced by caffeine intake.

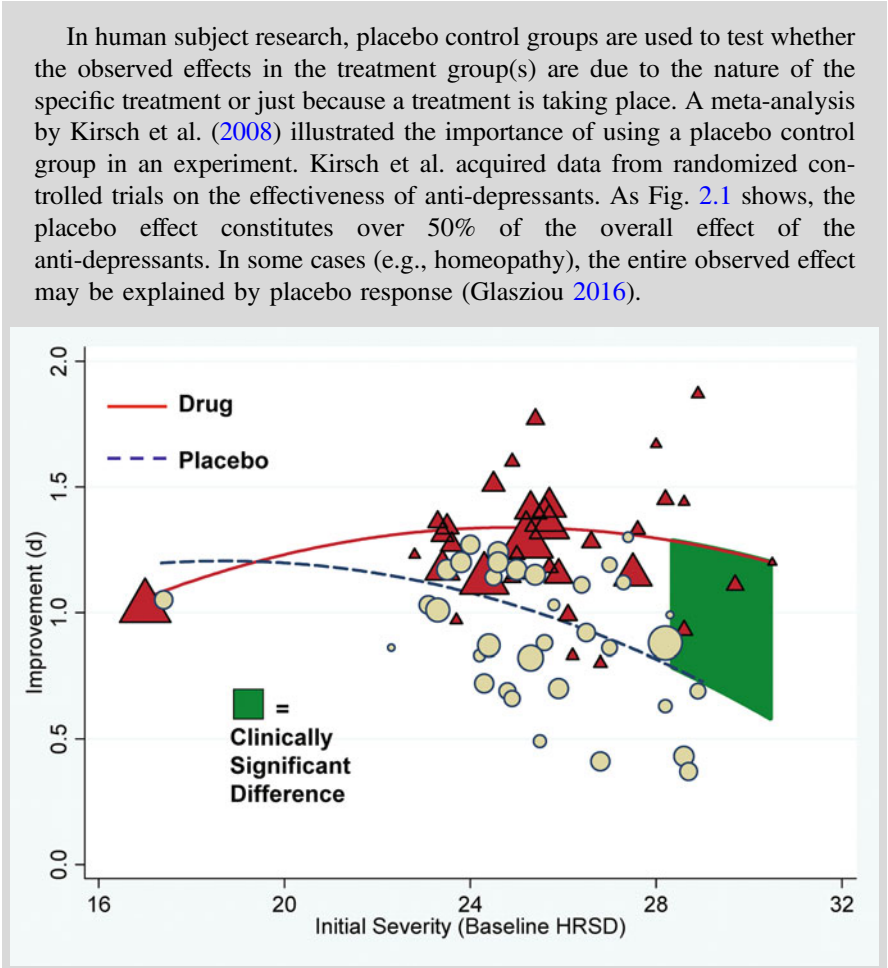


Fig. 2.1 The *horizontal axis* shows the participants' mean initial score on depression severity (*HRSD* = Hamilton Rating Scale for Depression; a score between 0 and 7 is generally considered normal, whereas a score of 23 or higher indicates very severe depression, with a maximum possible score of 52). The *vertical axis* shows the standardized mean difference (*d*, defined here as the change in *HRSD* divided by the standard deviation of the change). Thus, *d* is an effect size measure describing how much the symptoms improved with respect to the baseline score. *Triangles* = drug effect; *Circles* = placebo effect. The area of each *circle/triangle* is a function of the sample size. Sample sizes ranged between 10 and 403. Figure taken from Kirsch et al. (2008)

2.2.2 Participant Assignment to Groups

The assignment of participants to groups can be done randomly, a method known as simple randomization. Simple randomization can be done by means of a random number generator (e.g., in MATLAB, `round(rand(30,1))` produces a vector of zeros and ones).

Particularly when the sample size is small, simple randomization may lead to groups with unequal covariates, such as gender and age. Covariates are variables that, next to the variables that are manipulated (independent variables), are also predictors of the outcome (dependent) variables. To reduce imbalance between groups, minimization can be used. In minimization, participants are assigned to groups depending on the current group composition in order to minimize differences between the groups in terms of covariates (Pocock and Simon 1975; Taves 1974; for an overview of randomization techniques, such as block and stratified randomization, see Kang et al. 2008).

It is crucial that the assignment of participants to the experimental conditions is controlled by the researcher, because this prevents self-selection bias (i.e., participants choosing which group they are part of). It is not acceptable to test an experimental group and use data from a control group collected several months before. After all, conditions such as the outside weather, the quality of the measurement equipment, and the type of available participants may have changed in the meantime. In his lecture *Some remarks on science, pseudoscience, and learning how to not fool yourself*, Feynman referred to such lack of experimental rigor as “Cargo cult science” (Feynman 1974; see Textbox 2.2).

Textbox 2.2 Remarks by Richard Feynman on “Cargo cult science”

“When I was at Cornell, I often talked to the people in the psychology department. One of the students told me she wanted to do an experiment that went something like this—I don’t remember it in detail, but it had been found by others that under certain circumstances, X, rats did something, A. She was curious as to whether, if she changed the circumstances to Y, they would still do, A. So her proposal was to do the experiment under circumstances Y and see if they still did A.

I explained to her that it was necessary first to repeat in her laboratory the experiment of the other person—to do it under condition X to see if she could also get result A—and then change to Y and see if A changed. Then she would know that the real difference was the thing she thought she had under control. She was very delighted with this new idea, and went to her professor. And his reply was, no, you cannot do that, because the experiment has already been done and you would be wasting time.

Nowadays there’s a certain danger of the same thing happening, even in the famous field of physics. I was shocked to hear of an experiment done at the big accelerator at the National Accelerator Laboratory, where a person

used deuterium. In order to compare his heavy hydrogen results to what might happen to light hydrogen he had to use data from someone else's experiment on light hydrogen, which was done on different apparatus. When asked he said it was because he couldn't get time on the program (because there's so little time and it's such expensive apparatus) to do the experiment with light hydrogen on this apparatus because there wouldn't be any new result. And so the men in charge of programs at NAL are so anxious for new results, in order to get more money to keep the thing going for public relations purposes, they are destroying—possibly—the value of the experiments themselves, which is the whole purpose of the thing. It is often hard for the experimenters there to complete their work as their scientific integrity demands” (Feynman 1974, pp. 12–13; quoted with permission from *Engineering and Science*, published by California Institute of Technology).

2.3 Within-Subject Design

A between-subjects design has a major drawback: if the true effects are small, large sample sizes are needed to determine the existence of differences between the groups. For example, suppose one wants to test whether an in-vehicle warning system reduces the speed of car drivers. If the effect is small (a speed reduction of 3 km/h) and the spread among drivers is large (a standard deviation among participants of 10 km/h), then a large number of participants (352 in this specific case) is required to achieve an 80% probability of detecting the effect (i.e., a statistical power of 80%, as will be further explained in Sect. 3.3.2) for a Type I error rate of 5%. To achieve sufficient statistical power, randomized trials in the medical field may involve thousands participants and cost (hundreds of) millions of Euros (Biglan et al. 2000; Ioannidis 2013). Thus, a between-subjects experiment may not always be practically or financially feasible. Note that it is ethically problematic to run an underpowered experiment, because this means that resources and participants' time are wasted.

However, there is a solution: the within-subject design. In a within-subject design, also called repeated measures design, each participant undergoes multiple conditions. The advantage of a within-subject experiment is that the statistical power is usually higher than that of a between-subjects experiment, especially when the participants' scores for the different conditions are correlated, that is, when participants are consistent with respect to themselves across the different conditions. Because participants essentially serve as their own control, within-subject experiments require half as many participants as between-subjects experiments, or even considerably fewer, if the experimental conditions are positively correlated (Textbox 2.3).

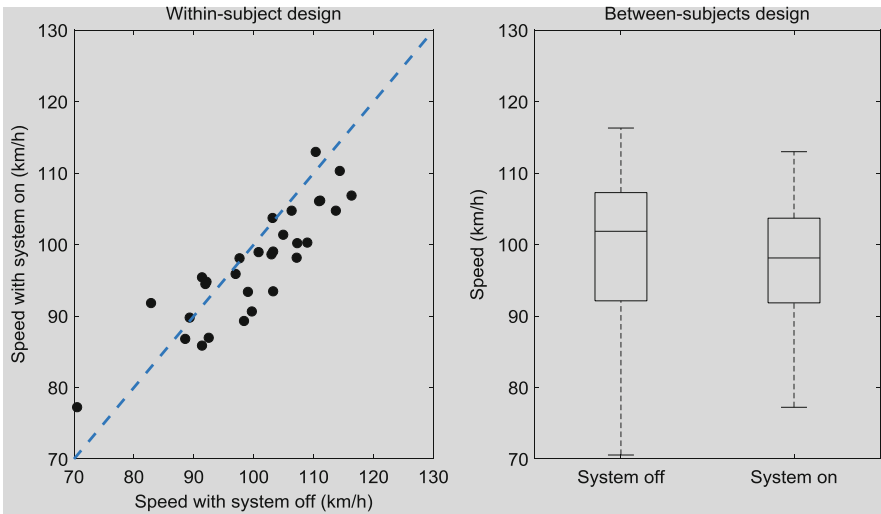


Fig. 2.2 Within-subject versus between-subjects design. In the within-subject experiment, a speed reduction of about 3 km/h can be reasonably well detected (by the naked eye in this case; Chap. 3 covers the corresponding statistical testing) at the level of individual participants. After all, 21 of the 30 participants drove slower with the in-vehicle warning system on than with the in-vehicle warning system off (i.e., the *dots* lie below the *line* of unity). The same effect cannot be reliably distinguished between groups, despite the fact that the sample size is twice as large as in the within-subject design. In both figures, the means (M) and standard deviations (SD) are as follows: $M_{off} = 100.22$ km/h, $M_{on} = 97.20$ km/h, $SD_{off} = 10.26$ km/h, $SD_{on} = 7.99$ km/h. Correlation coefficient between the speed with the system on and off = 0.88. These values were drawn from a population with means of 100 and 97 km/h, respectively, a standard deviation per group of 10 km/h, and a correlation of 0.90

Textbox 2.3 Illustration of the statistical power of a within-subject versus a between-subjects experiment

Suppose that a researcher aims to investigate whether an in-vehicle warning system reduces the speed of car drivers. Figure 2.2 provides simulated results of a within-subject experiment with 30 participants, each of whom drove with the in-vehicle warning system on and off, and of a between-subjects experiment with 60 participants (30 per group: one group driving with the in-vehicle warning system on and another group driving with the in-vehicle warning system off). The speed-reduction effect is easier to detect in the within-subject experiment.

Because of its higher statistical power, a within-subject experiment is often preferred in human subject research in engineering. However, there are important drawbacks: in a within-subject experiment, each participant encounters each experimental condition and is therefore likely to be occupied for a longer time than

in a between-subjects experiment. Moreover, the results are susceptible to order effects, such as practice and fatigue, and to carryover effects. A solution to these problems is counterbalancing.

Counterbalancing is a method that aims to control for order and carryover effects by letting participants undergo the various conditions in different orders. There are several approaches to defining these orders. One can counterbalance with all possible orders (Underwood 1949), also called complete counterbalancing. For example, for four conditions, there are $4! = 24$ possible orders (in MATLAB: `n=4 ; perms(1:n)`). This also means that at least 24 participants are needed for this experiment. Note, however, that the number of possible orders increases rapidly, with five conditions leading to 120, and six conditions leading to 720 permutations. A technique that generates a workable number of possible orders is the Latin square. In a Latin square, each of the n conditions appears exactly one time in each row and exactly one time in each column. In MATLAB, `n=4 ; M=[1:n; ones(n-1,n)] ; M=rem(cumsum(M)-1,n)+1 ;` generates a Latin square for four conditions (Van der Geest 2009a).

1	2	3	4
2	3	4	1
3	4	1	2
4	1	2	3

Here, the columns represent the order in which the condition number is presented to the participant, and the rows represent the participant number. In other words, each of the four conditions (1, 2, 3, or 4) is encountered once in the first session, once in the second session, once in the third session, and once in the fourth session.

A drawback of the Latin square shown above is that all conditions are surrounded by the same pattern of conditions throughout the Latin square, which means that not all order and carryover effects can be ruled out. Williams (1949) defined complete Latin squares, in which for each element i , element j immediately follows element i exactly once (for a MATLAB function, see Van der Geest 2009b). For example:

1	2	4	3
2	3	1	4
3	4	2	1
4	1	3	2

The disadvantage of a complete Latin square is that it cannot be defined for all odd numbers of conditions. Campbell and Geller (1980) introduced so-called balanced Latin squares, which can also be computed if n is odd. In a balanced Latin square, each pair of conditions i and j appears exactly twice adjacent to each other. An example of a balanced Latin square for $n = 5$ is shown below:

1	2	3	4	5
2	4	1	5	3
3	1	5	2	4
4	5	2	3	1
5	3	4	1	2

For more advanced types of balanced Latin squares, the reader is referred to Alimena (1962), Colbourn et al. (1996), and Kim and Kim (2010).

When conducting a within-subject experiment, it is of paramount importance to apply counterbalancing. Without counterbalancing, the results are likely to be invalid. Referring to the example in Textbox 2.3, it is known that drivers tend to drive slightly faster at the end of an experiment as compared to the beginning (e.g., Mars et al. 2014). Thus, if all participants drive first with the in-vehicle warning system off and then with the system on (or vice versa), it becomes impossible to separate the treatment effect (i.e., the effect of the in-vehicle warning system) from the practice effect. Note that although counterbalancing is a powerful technique, it rests on the assumption that the degree of practice, fatigue, and carryover effects do not interact with the treatment. For example, if driving with the in-vehicle system on yields a steeper learning curve than driving with the system off, then the above-presented results are not perfectly valid, despite counterbalancing.

Carryover effects could be reduced by implementing training or practice sessions prior to the start of the experiment. This means that the learning curve has flattened out so that practice effects are reduced during the actual experiment (Greenwald 1976). It is also possible to train participants to proficiency (e.g., conducting per participant as many practice sessions as needed in order to reach a target score) prior to the experiment. Note, however, that when participants are trained to proficiency, the outcome variables are confounded by the training time. One can also reduce carryover effects in a counterbalanced design by using large time intervals between conditions, for example by letting each participant complete each condition on a different day (Greenwald 1976; Keren 1993).

2.4 Choosing Between-Subjects or Within-Subject Design: More Than just a Matter of Statistical Power

Whether one chooses a between-subjects or a within-subject design is not just a matter of statistical power. As explained above, within-subject studies have the disadvantage of order and carryover effects. Moreover, there are theoretical considerations to be taken into account: whether one wants to obtain knowledge about average performance of groups (as is acquired with a between-subjects experiment) or about whether individuals get higher scores in condition A than in condition B (as is acquired with a within-subject experiment). The results of between-subjects

In some cases, a between-subjects design is the only way to test a particular hypothesis. If one is interested in the effect of training on the participants' performance or behaviour, a between-subjects design rather than a within-subject design should be used. For example, in order to test whether an online course leads to higher student grades than a course of similar content taught in the lecture room, a between-subjects experiment should be conducted. A within-subject experiment is not possible in this case, because a person can learn a course only once. Because of the large samples that are required for between-subjects research, as well as the logistics and ethical difficulties involved (e.g., whether it is acceptable to educate different student groups with educational methods that are expected to have different degrees of effectiveness; for an overview of such ethical concerns, see Borman 2002; Burtless 2002), experiments in educational research are rare (Thompson et al. 2005).

Birnbaum (1999) conducted an online experiment in which he asked participants to judge either how large the number 221 is or how large the number 9 is, on a 10-point scale where 1 = *very very small* and 10 = *very very large* (Fig. 2.3). 45 and 40 people completed the 9- and 221-number judgement experiments, respectively.

The mean judgement was significantly higher for the number 9 than for the number 221 (mean for the number 9 = 5.13 vs. mean for the number 221 = 3.10). This counterintuitive result can be explained by the fact that the experiment was conducted between subjects, meaning that each participant was presented with only one of the numbers.

Fig. 2.3 Online judgment experiment conducted by Birnbaum (1999). Screenshots taken from <http://psych.fullerton.edu/mbirnbaum/done.htm> (11 November 2016) with permission from Prof. M. Birnbaum

2.5 Validity Threats in Experiments

2.5.1 Demand Characteristics

Experimental validity refers to whether what is investigated represents what was supposed to be investigated. An important threat to the validity of an experiment is *demand characteristics*, which refers to participants' tendency to adjust their behaviour according to what they think that the experimenter expects from them. For example, suppose that an experiment is conducted to investigate the user-friendliness of a computer program, and that a participant figures out that one of the computer programs (A) has been developed by the experimenter himself. In this case, the participant may think that the experimenter expects computer program A to be user-friendlier than computer program B and thus may (unconsciously) try harder and achieve a better performance when working with program A as compared to when working with program B. The placebo effect presented in Fig. 2.1 illustrates how strong the impact of expectancies may be.

Within-subject experiments are more susceptible to demand characteristics than between-subjects experiments. In a within-subject experiment, the participant undergoes multiple experimental conditions, which makes the participant easily aware of the differences between these conditions (Charness et al. 2012). A technique to protect an experiment from demand characteristics is *blinding* (also called *masking*), which means that the experimenter does not disclose which experimental condition the participant receives.

2.5.2 Experimenter Expectancy Effect

The validity of an experiment may also be compromised by the experimenters themselves. An experimenter may hold expectations regarding his/her own hypothesis and may (unconsciously) express enthusiasm that influences the participants' behaviour (see Rosenthal et al. 1966 for experiments showing how the experimenter's talking speed, hand gestures, and facial expressions may influence the experimental results). Although the experimenters should of course answer any questions the participants may have (see also the topic of informed consent treated in Sect. 1.4.1), it is important that experimenters do not engage in lengthy conversations with the participants and remain neutral during the experiment. Experimenter expectancy bias is not only relevant in human subject research, but also in physical sciences and engineering (see Textbox 2.5 for an example).

Textbox 2.5 Experimenter expectancy bias in physics

A classic example of experimenter expectancy bias in physics is that of imaginary N-rays observed by several researchers at the beginning of the 20th century. In 1902, physicist Prosper-René Blondlot claimed the discovery of a new type of radiation, N-rays, emitted by a variety of metals and increasing the luminosity of white surfaces in a dark room. In 1904, 77 papers on N-rays were listed in *Science Abstracts* (Ederer 1975).

In 1905, however, Pozdëna conducted a double-blinded experiment that disproved the existence of N-rays. Moreover, Wood (1904) reported that in an attempt to test the credibility of Blondlot's observations, he secretly replaced the metal surface that was supposed to emit the rays with a wooden surface; nevertheless, Blondlot still argued that luminosity increased in a dark room, a result that pointed towards experimenter expectancy bias. After the reports by Pozdëna and Wood, the number of papers claiming that N-rays exist reduced considerably, with *Science Abstracts* counting only eight papers on N-rays in 1905 and zero after 1909 (Ederer 1975).

A technique to protect an experiment from experimenter expectancy bias is *double blinding*, meaning that neither the participant nor the research team are aware of which treatment the participant receives. In the medical field, double blinding is common and relatively easy to achieve. After all, a placebo pill can be prepared in such a way that it looks, smells, tastes, and even has similar physiological side effects (e.g., change in urine colour; Stoney and Johnson 2012) as a real pill (although making real and placebo pills identical has not always been successful; see Friedman et al. 2015). In typical human subject research in engineering, blinding may be difficult to achieve, because the stimuli or tasks are easily distinguishable by both the participant and the experimenter. For example, it is easy to distinguish different types of feedback (e.g., audio vs. visual) or the state of a device (e.g., a motion platform being on or off).

2.6 Quasi-experiments

Sometimes it is undesirable or impossible to conduct an experiment. For example, when concerns started to arise in the 1940s that smoking may cause lung cancer, researchers faced constraints. Clearly, it is not ethically or practically feasible to conduct an experiment in which 50% of participants are assigned to a 'smoking group' with the instruction to smoke a number of cigarettes per day for several decades, whereas the remaining 50% are assigned to a control group not allowed to smoke. In this case, quasi-experiments had to be conducted instead.

A variety of quasi-experimental studies on the health risk of smoking have indeed been conducted over the last few decades. Examples are cohort studies, in

which smoking and non-smoking individuals were followed for several years to investigate lung cancer rates (e.g., Freedman et al. 2008), and case-control studies, in which individuals with and without lung cancer were recruited and their smoking history was investigated (e.g., Peto et al. 2000). Additionally, a large body of knowledge has been gathered on the biological effects of cigarette smoke compounds on bonding with DNA and on associated genetic mutations (Centers for Disease Control and Prevention 2010). Based on such quasi-experimental studies, epidemiologists are now able to conclude that smoking is the single greatest cause of preventable death worldwide (U.S. Department of Health and Human Services 2014) and that people who have been smoking since youth die on average 10 years younger than those who have never smoked (Doll et al. 2004).

Cohort studies are studies in which individuals with a common baseline characteristic (e.g., age) are analysed. The researchers identify the individuals within the cohort who have been exposed to a risk factor (e.g., smoking) the outcome of which is of interest (e.g., lung cancer), and compare these individuals with respect to the presence of this outcome to the members of the same cohort who have not been exposed to the risk factor. For example, in a prospective cohort study investigating the long-term effects of monocular head-mounted displays (risk factor) on visual complaints (outcome variable), researchers selected a cohort of pilots of similar ages, half of which were serving as Apache army pilots (who typically use monocular displays) and the other half serving as non-Apache helicopter pilots (therefore not using monocular displays) and compared the visual complaints of the two groups annually for a period of 10 years (Hiatt et al. 2001).

In case-control studies, researchers identify two groups that differ in an outcome variable and compare their characteristics in terms of a risk factor that is expected to have an effect on the outcome variable. For example, in a study investigating whether not wearing a bicycle safety helmet (risk factor) is associated with a higher rate of head injuries in a bicycling accident (outcome variable) as compared to wearing a helmet, researchers identified a group with and a group without head injury in a bicycling accident and compared the proportion of individuals within each group wearing a bicycle safety helmet during the accident (Thompson et al. 1989). Figure 2.4 illustrates the difference between cohort and case-control studies.

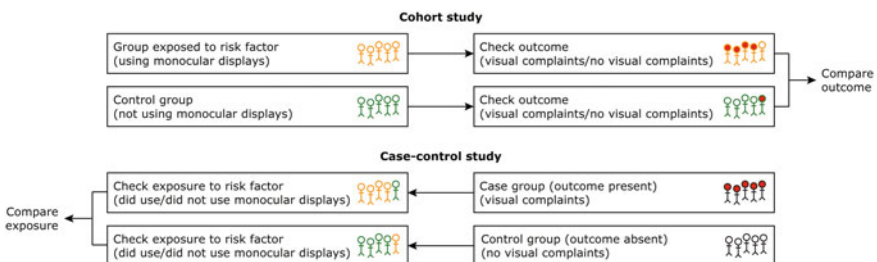


Fig. 2.4 Cohort study and case-control study design. *Orange* indicates participants exposed to the risk factor, and *green* indicates participants not exposed to the risk factor. *Red* annotates the participants who tested positive on the outcome variable

2.7 Validity Threat in Quasi-experiments: Causation Versus Correlation

It is a challenge to draw causal inferences from a quasi-experimental design, because the hypothesized causal relationship between the risk factor and the outcome variable might be confounded. A confounder is a variable that relates to both the risk factor and the outcome variable of a study while not being part of the causal pathway between the risk factor and the outcome variable (that is, a confounder is a common cause of both variables). For example, in a study investigating the effect of wearing a helmet on the risk of head injury, a confounder might be that cyclists not wearing helmets are also less likely to use lights when riding in the dark as compared to cyclists who wear a helmet (McGuire and Smith 2000), thereby being at a risk of suffering an injury of higher severity than cyclists using lights (Wang et al. 2015). Despite the risk that the causal pathway might be confounded, a quasi-experimental design can still lead to causal interpretations, by controlling for said confounders. For example, in a case-control study, if age is considered to be a confounder, cases and controls can be matched with respect to age. A statistical model, such as regression analysis, can also be used to control for confounders.

According to the Bradford Hill's criteria of causality, causality can be distinguished from mere association based on the following nine principles: (1) strength, (2) consistency, (3) specificity, (4) temporality, (5) biological gradient, (6) plausibility, (7) coherence, (8) experiment, and (9) analogy (Hill 1965). For example, in the case of smoking and lung cancer, there is evidence regarding a biological gradient (criterion 5), in the sense that a dose-response relationship exists: the earlier one quits smoking and the fewer cigarettes one smokes per day, the greater one's expected lifespan (Doll et al. 2004). There is also evidence regarding temporality (criterion 4), namely a 20–25-year lag between nationwide trends in cigarette smoking and the incidence of lung cancer (Peto et al. 2000; Shibuya et al. 2005). Moreover, the effect of smoking on lung cancer is strong (criterion 1). In the case of smoking, it is thus possible to rule out the effect of confounders with relative ease. Many other quasi-experimental studies, on the other hand, do suffer from confounding variables and small effects. For example, the effect of meat consumption and cancer risk remains controversial, with smoking and physical activity being important confounders (e.g., Sinha et al. 2009; for more examples of confounders of the relationship between diet and cancer, see Key et al. 2002).

2.8 Validity Threats in Experiments and Quasi-experiments

2.8.1 Sample Representativeness

Eligible participants can be recruited via flyers, by e-mail, or by contacting students or peers. When sampling participants, it is important to reflect on the

representativeness of the sample in relation to the hypothesis (for an overview of sampling techniques, see Henry 1990). Henrich et al. (2010) argued that many of the published research findings may not be generalizable, because participants are almost always sampled from ‘WEIRD’ (Western, Educated, Industrialized, Rich, and Democratic) populations. On the other hand, in a ‘many labs’ study with 6344 participants taking part in various psychological experiments, it was found that the effect sizes were similar regardless of whether the experiment was done in a lab or via the Internet, and regardless of whether the experiment was done in or outside the United States (Klein et al. 2014). A specific issue at technical universities is that engineering students have above-average spatial skills (Wai et al. 2009) and that males are over-represented. For example, about 80% of the students at the Delft University of Technology are male (De Winter and Dodou 2011; see also Van Leeuwen et al. 2014, in which only 14 out of the 62 participants recruited from the student community of the Delft University of Technology were females). This means that the measured effects may not hold for the general population.

2.8.2 Attrition

Attrition refers to a decline of the number of participants over the course of a study. Attrition is common in cohort studies but also in experiments consisting of several phases, as participants may not return for the follow-up. Attrition becomes a threat to the validity of the study especially when attrition is imbalanced between participant groups or conditions (e.g., when the treatment group loses more participants than the control group; a phenomenon called *selective attrition* or *differential attrition*). Attrition is also problematic when the type of participants quitting the experiment differs from the returning participants.

2.9 Measurements and Measures

After the experiment has been completed, the data are usually submitted to a statistical test (for more information, see Chap. 3). When setting up an experiment, it is important to distinguish between a measurement and a measure. One may, for example, perform a *measurement* of the speed of a car at a sampling frequency of 100 Hz. Based on this measurement it is possible to define a *measure*, such as the mean speed during a trial. After the experiment has been completed, the *measures* rather than the *measurements* are subjected to a statistical test, for example to test the hypothesis that using an in-vehicle device reduces speed compared to not using the in-vehicle device. Measures need to be operationalized so that they are reproducible. Instead of writing: ‘*The outcome measure was the speed of the car*’, it is better to write: ‘*The outcome measure was the mean speed of the car from the moment the car entered the highway until the moment the car left the highway (km/h).*’

2.9.1 Statistical Reliability

Reliability refers to the repeatability (i.e., consistency) of the measurements. In engineering, measurements are usually highly reliable, especially when the environmental conditions are kept constant. Put simply, the measured weight, length, or mass of an object remain almost constant over a series of measurement repetitions (see also Sect. 1.1).

In human subject research, on the other hand, there is often a high degree of noise in the measurement, because humans exhibit moment-to-moment variation. For example, the reliability coefficient (i.e., the test-retest correlation) of a single reaction time to a stimulus is only about 0.20 (Jensen 2006; Johnson et al. 1985). If the reliability coefficient is low, the correlation coefficient with an external variable is low as well, and so is the statistical power (Liu and Salvendy 2009; Rushton et al. 1983; Schmidt and Hunter 1999).

A single measurement of a human participant (a single item on a questionnaire, a single reaction time, a single speed measurement of a car) is statistically unreliable and therefore of limited use. Reliability can be improved by averaging across multiple measurement instances (see Textbox 2.6). For example, in order to obtain a reliability coefficient of 0.90 of a person's reaction time to a visual stimulus, and assuming a reliability (i.e., test-retest correlation) of 0.20 for single reaction time measurements, 36 trials of this person need to be averaged [calculated using Eq. (2.1)]. Similarly, it is advisable to calculate a total/average score across multiple questionnaire items rather than to use a single questionnaire item, and it is wise to measure the average speed of a car along a road segment rather than to rely on a single speed trap.

Textbox 2.6 Weight judgement: Wisdom of the crowd

In an experiment, Gordon (1924) used 10 weights of similar appearance, ranging from 16 to 17.6 g, with equal increments between weights. She then asked 200 participants to sort the weights in decreasing order. The correlation between the order proposed by each participant and the true order of weights was then calculated. The 200 correlations were found to vary greatly, between +0.95 and -0.81, with a mean of +0.41.

Next, Gordon clustered the participant judgements into groups. For each group she calculated an 'average order' by taking the average of the positions assigned to each weight. The correlations between these 'average orders' and the true order were calculated as a function of the size of the group (Fig. 2.5):

Mean of 40 groups of 5 participants = 0.68
 Mean of 20 groups of 10 participants = 0.79
 Mean of 10 groups of 20 participants = 0.86
 Mean of 4 groups of 50 participants = 0.94

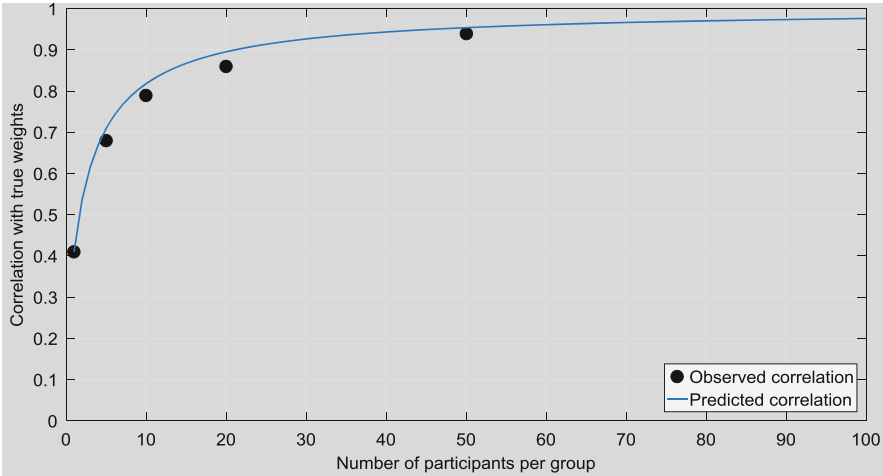


Fig. 2.5 Mean correlation coefficient between the group’s ‘average order’ and the true values of the weights as a function of the number of participants in the group (Gordon 1924). Figure based on Eysenck (1939)

The Spearman-Brown prediction formula (Eq. (2.1)) says that the reliability of a test (R) improves according to the number of combined tests (n) and the reliability of the current test (r). This formula assumes that the measurement errors are independent. In the case of the abovementioned weight-ordering experiment, r equals 0.41^2 .

$$R = \frac{nr}{1 + (n - 1)r} \quad (2.1)$$

2.9.2 Self-reported Measures (Questionnaires)

In human subject research, self-reported data can be collected to complement data measured by sensors and instruments. The strength of questionnaires is that they allow a researcher to gain insight into private characteristics and states that cannot be directly observed, such as opinion, strategy, and experienced workload. Moreover, self-reported data can offer insight into the history of the participants (e.g., gaming experience, daily habits). Questionnaires can add much to the research and aid in interpreting the data measured by sensors and instruments.

Preparing a new questionnaire is a challenging task, and the validity of the self-reported data depends on the way the questions are formulated. As Schwarz (1999) put it: “Questions shape the answers” (see Textbox 2.7 for an example).

Generally, however, there is no need to make your own questionnaires: there are thousands of validated questionnaires in the scientific literature, which you could use (but do not forget to acknowledge the original authors by citing them; more of which will be discussed in Sect. 4.2.3). If a new questionnaire needs to be created, several decisions have to be made concerning the type of questions (open-ended, closed-ended, or a combination of both), mode of administration (paper-and-pencil, computer, online), number of items, number of response options, type of labelling of the response options, etc. For guidelines on how to set up a questionnaire, see Krosnick and Presser (2010).

Textbox 2.7 Self-reports: how the question shapes the answer

Schwarz (1999) provided examples that illustrate how responses may differ, depending on the formulation of a question and response options. For example, when parents were asked: “*What is the most important thing for children to prepare them for life?*”, the response “*To think for themselves*” was given by 61.5% of the parents when this option was provided in a list of response options and by 4.6% of the parents when the question was open and no response options were provided.

Loftus and Palmer (1974) conducted an experiment in which participants watched films with car accidents and then responded to questions regarding these films. Participants who were asked “*About how fast were the cars going when they smashed into each other?*” provided higher estimates of speed than participants who were asked the same question but with “*collided*”, “*bumped*”, “*contacted*”, or “*hit*” instead of “*smashed*”. Moreover, one week after the viewing of the films, participants who were asked the abovementioned question with the verb “*smashed*” were more likely to give a positive response to the question “*Did you see any broken glass?*”, despite the fact that broken glass was not visible in the films.

Self-report questionnaires may suffer from several pitfalls:

- *Social desirability*. Social desirability is the tendency of participants to provide answers that are socially acceptable and to be more reluctant to disclose embarrassing facts about themselves. Social desirability bias can be somewhat reduced by making questionnaires anonymous (Buchanan 2000; Dodou and De Winter 2014). To detect social desirability, the questionnaire can include a so-called lie scale, that is, a set of questions about human foibles and moral weaknesses (e.g., “*Are all your habits good and desirable ones?*”, “*Have you ever cheated at a game?*”; Eysenck et al. 1985). Strong agreement (or disagreement in the case of negative statements) with such statements indicates a tendency to present oneself in a socially desirable manner.

- ‘Above-average effect’ or ‘illusory superiority’ (Kruger and Dunning 1999). This is the tendency of humans to report that they perform better than the average. For example, most people report higher frequencies of healthy behaviours for themselves than they report for the average other (Hoorens and Harris 1998).
 - *Response style* (Jackson and Messick 1958; for an overview, see Van Vaerenbergh and Thomas 2013). Response style refers to the tendency/bias to respond in a similar manner regardless of the content of the question. There are several types of response styles:
 - *Extreme response style* refers to a tendency to give extremely low or high responses on rating scales.
 - *Moderacy bias* (or *mid-point response style*) is the inverse tendency of extreme response style, namely to give medium ratings for all questions; a similar bias is the *mild-response style*, in which the extremely high or low response options are avoided.
 - *Yea-saying* (or *acquiescence bias*) and *nay-saying* (*dis-acquiescence*) refers to a tendency of a participant to give overly positive and overly negative responses, respectively. This is different from extreme response style, where both extremely positive and extremely negative responses are present. Acquiescence and dis-acquiescence can lead to spurious correlations. Take a two-item questionnaire, each item consisting of a five-point Likert scale (1 = *strongly disagree*, 5 = *strongly agree*). If 5% of the participants spuriously tick ‘*strongly disagree*’ on both items, and another 5% of the participants spuriously tick ‘*strongly agree*’ on both items, a moderate correlation of $r = 0.18$ arises between the two items, while the expected correlation is $r = 0.00$. This is demonstrated with the following MATLAB simulation:
- ```
A=ceil(5*rand(1000000,2)); % random responses of 1,000,000
participants for two items (1 = strongly disagree, 5 =
strongly agree)
r1=corr(A(:,1),A(:,2)); % correlation coefficient between
the two items
A(1:50000,:)=1; % 50,000 participants ticked 'strongly
disagree' on both items
A(50001:100000,:)=5; % 50,000 participants ticked 'strong-
ly agree' on both items
r2=corr(A(:,1),A(:,2)); % correlation coefficient between
the two items
disp([r1 r2])
```
- *Anchoring* is the tendency to use the first response as reference for the rest of the responses in a questionnaire, even when the questions are unrelated and therefore using one of the responses as reference is meaningless (Tversky and Kahneman 1974).

A common remedy against response style is to reverse the order of response options for some items (e.g., instead of ordering the response options from ‘*strongly agree*’ to ‘*strongly disagree*’, one can order the response options for some items

from ‘*strongly disagree*’ to ‘*strongly agree*’). Another remedy is temporal separation of questionnaires (i.e., questionnaires completed on different days). Moreover, as mentioned above, it is advisable to use self-reported questionnaires in conjunction with data recorded by sensors and other measurement equipment.

## 2.10 Finally, Some Tips Before Starting the Experiment

- *Perform a pilot study.* A pilot study refers to preliminary measurements conducted in order to evaluate, among other things, the feasibility and soundness of the experimental design, the safety of the procedures, the clarity of the participant instructions, as well as time- and administration-related bottlenecks. Furthermore, a pilot study allows for estimating the expected effect sizes and required sample sizes.
- *Check the measurement equipment.* Before starting an experiment, it is important to check whether sensors and other measurement equipment have been positioned at the right locations and function well (e.g., free of noise or vibrations), to calibrate instruments, and to rule out interferences (e.g., electromagnetic interferences).
- *Prepare the experiment carefully.* It is important to remember that preparing an experiment occupies a large proportion of time in a research project. The execution of the experiment itself may take less than a few days, and is usually straightforward if the experiment is well prepared.

## References

- Alimena, B. S. (1962). A method of determining unbiased distribution in the Latin square. *Psychometrika*, 27, 315–317. <https://doi.org/10.1007/BF02289627>
- Bhatt, A. (2010). Evolution of clinical research: A history before and beyond James Lind. *Perspectives in Clinical Research*, 1, 6–10.
- Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science*, 1, 31–49. <https://doi.org/10.1023/A:1010024016308>
- Birnbaum, M. H. (1999). How to show that  $9 > 221$ : Collect judgments in a between-subjects design. *Psychological Methods*, 4, 243–249. <https://doi.org/10.1037/1082-989X.4.3.243>
- Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education*, 77, 7–27. [https://doi.org/10.1207/S15327930PJE7704\\_2](https://doi.org/10.1207/S15327930PJE7704_2)
- Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121–140). San Diego, CA: Academic Press.
- Burtless, G. (2002). Randomized field trials for policy evaluation: Why not in education? In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 179–197). Washington, DC: Brookings Institute.

- Campbell, G., & Geller, S. (1980). Balanced Latin squares (Mimeoseries No. 80-26). West Lafayette, IN: Department of Statistics, Purdue University.
- Centers for Disease Control and Prevention. (2010). *How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general*. Centers for Disease Control and Prevention (US).
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81, 1–8. <https://doi.org/10.1016/j.jebo.2011.08.009>
- Colbourn, C. J., Dinitz, J. H., & Wanless, I. M. (1996). Latin squares. In C. J. Colbourn & J. H. Dinitz (Eds.), *Handbook of combinatorial designs* (pp. 135–152). Boca Raton, FL: CRC Press, Taylor & Francis Group.
- Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276, 889–897. <https://doi.org/10.1001/jama.1996.03540110043030>
- De Craen, A. J., Kaptchuk, T. J., Tijssen, J. G., & Kleijnen, J. (1999). Placebos and placebo effects in medicine: Historical overview. *Journal of the Royal Society of Medicine*, 92, 511–515.
- De Winter, J. C. F., & Dodou, D. (2011). Predicting academic performance in engineering using high school exam scores. *International Journal of Engineering Education*, 27, 1343–1351.
- Dodou, D., & De Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, 36, 487–495. <https://doi.org/10.1016/j.chb.2014.04.005>
- Doll, R., Peto, R., Boreham, J., & Sutherland, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ*, 328, 1519. <https://doi.org/10.1136/bmj.38142.554479.AE>
- Ederer, F. (1975). Patient bias, investigator bias and the double-masked procedure in clinical trials. *The American Journal of Medicine*, 58, 295–299. [https://doi.org/10.1016/0002-9343\(75\)90594-X](https://doi.org/10.1016/0002-9343(75)90594-X)
- Eysenck, H. J. (1939). The validity of judgments as a function of the number of judges. *Journal of Experimental Psychology*, 25, 650–654. <https://doi.org/10.1037/h0058754>
- Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6, 21–29. [https://doi.org/10.1016/0191-8869\(85\)90026-1](https://doi.org/10.1016/0191-8869(85)90026-1)
- Feynman, R. P. (1974). Cargo cult science. Some remarks on science, pseudoscience, and learning how to not fool yourself. Caltech's 1974 commencement address. <http://calteches.library.caltech.edu/51/2/CargoCult.pdf>
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance*, 23, 339–359. [https://doi.org/10.1016/0030-5073\(79\)90002-3](https://doi.org/10.1016/0030-5073(79)90002-3)
- Freedman, N. D., Leitzmann, M. F., Hollenbeck, A. R., Schatzkin, A., & Abnet, C. C. (2008). Cigarette smoking and subsequent risk of lung cancer in men and women: Analysis of a prospective cohort study. *The Lancet Oncology*, 9, 649–656. [https://doi.org/10.1016/S1470-2045\(08\)70154-2](https://doi.org/10.1016/S1470-2045(08)70154-2)
- Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (2015). *Fundamentals of clinical trials* (5th ed.). Springer International Publishing.
- Glasziou, P. (2016, February 16). Still no evidence for homeopathy [blog]. *The BMJ blogs*. <http://blogs.bmj.com/bmj/2016/02/16/paul-glasziou-still-no-evidence-for-homeopathy/>
- Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, 7, 398–400. <https://doi.org/10.1037/h0074666>
- Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83, 314–320. <https://doi.org/10.1037/0033-2909.83.2.314>
- Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., et al. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 366, 924–926. <https://doi.org/10.1136/bmj.39489.470347.AD>

- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Henry, G. T. (1990). *Practical sampling*. Newbury Park, CA: Sage Publications.
- Hiatt, K. L., Braithwaite, M. G., Crowley, J. S., Rash, C. E., Van de Pol, C., Ranchino, D. J., et al. (2001). *The effect of a monocular helmet-mounted display on aircrew health: A cohort study of Apache AH MK1 pilots* (Initial Report No. USAARL-2002-04). Fort Rucker, AL: U.S. Army Aeromedical Research Laboratory.
- Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295–300.
- Hoorens, V., & Harris, P. (1998). Distortions in reports of health behaviors: The time span effect and illusory superiority. *Psychology and Health*, 13, 451–466. <https://doi.org/10.1080/08870449808407303>
- Ioannidis, J. P. (2013). Mega-trials for blockbusters. *JAMA*, 309, 239–240. <https://doi.org/10.1001/jama.2012.168095>
- Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin*, 55, 243–252. <https://doi.org/10.1037/h0045996>
- Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam: Elsevier.
- Johnson, R. C., McClearn, G. E., Yuen, S., Nagoshi, C. T., Ahern, F. M., & Cole, R. E. (1985). Galton's data a century later. *American Psychologist*, 40, 875–892. <https://doi.org/10.1037/0003-066X.40.8.875>
- Kaasinen, V., Aalto, S., Nägren, K., & Rinne, J. O. (2004). Expectation of caffeine induces dopaminergic responses in humans. *European Journal of Neuroscience*, 19, 2352–2356. <https://doi.org/10.1111/j.1460-9568.2004.03310.x>
- Kang, M., Ragan, B. G., & Park, J. H. (2008). Issues in outcomes research: An overview of randomization techniques for clinical trials. *Journal of Athletic Training*, 43, 215–221.
- Keren, G. (1993). Between-or within-subjects design: A methodological dilemma. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 257–272). Hillsdale, NJ: Erlbaum.
- Key, T. J., Allen, N. E., Spencer, E. A., & Travis, R. C. (2002). The effect of diet on risk of cancer. *The Lancet*, 360, 861–868. [https://doi.org/10.1016/S0140-6736\(02\)09958-0](https://doi.org/10.1016/S0140-6736(02)09958-0)
- Kim, B. G., & Kim, T. (2010). A program for making completely balanced Latin square designs employing a systemic method. *Revista Colombiana de Ciencias Pecuarias*, 23, 277–282.
- Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLOS Medicine*, 5, e45. <https://doi.org/10.1371/journal.pmed.0050045>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45, 142–152. <https://doi.org/10.1027/1864-9335/a000178>
- Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). West Yorkshire, England: Emerald Group.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Liu, Y., & Salvendy, G. (2009). Effects of measurement errors on psychometric measurements in ergonomics studies: Implications for correlations, ANOVA, linear regression, factor analysis, and linear discriminant analysis. *Ergonomics*, 52, 499–511. <https://doi.org/10.1080/00140130802392999>
- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13, 585–589. [https://doi.org/10.1016/S0022-5371\(74\)80011-3](https://doi.org/10.1016/S0022-5371(74)80011-3)

- Mars, F., Deroo, M., & Charron, C. (2014). Driver adaptation to haptic shared control of the steering wheel. In *Proceedings of the 2014 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1505–1509). <https://doi.org/10.1109/SMC.2014.6974129>
- McGuire, L., & Smith, N. (2000). Cycling safety: Injury prevention in Oxford cyclists. *Injury Prevention*, 6, 285–287. <https://doi.org/10.1136/ip.6.4.285>
- Mikalsen, A., Bertelsen, B., & Flaten, M. (2001). Effects of caffeine, caffeine-associated stimuli, and caffeine-related information on physiological and psychological arousal. *Psychopharmacology*, 157, 373–380. <https://doi.org/10.1007/s002130100841>
- National Health and Medical Research Council. (1999). *A guide to the development, implementation and evaluation of clinical practice guidelines*. Canberra: National Health and Medical Research Council.
- Oken, B. S. (2008). Placebo effects: Clinical aspects and neurobiology. *Brain*, 131, 2812–2823. <https://doi.org/10.1093/brain/awn116>
- Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., & Doll, R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: Combination of national statistics with two case-control studies. *BMJ*, 321, 323–329. <https://doi.org/10.1136/bmj.321.7257.323>
- Pocock, S. J., & Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31, 103–115. <https://doi.org/10.2307/2529712>
- Pozděna, R. F. (1905). Versuche über Blondlots „Emission pesante”. *Annalen der Physik*, 322, 104–131. <https://doi.org/10.1002/andp.19053220606>
- Rosenthal, R., Kohn, P., Greenfield, P. M., & Carota, N. (1966). Data desirability, experimenter expectancy, and the results of psychological research. *Journal of Personality and Social Psychology*, 3, 20–27. <https://doi.org/10.1037/h0022604>
- Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, 94, 18–38. <https://doi.org/10.1037/0033-2909.94.1.18>
- Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, 27, 183–198. [https://doi.org/10.1016/S0160-2896\(99\)00024-0](https://doi.org/10.1016/S0160-2896(99)00024-0)
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54, 93–105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Shibuya, K., Inoue, M., & Lopez, A. D. (2005). Statistical modeling and projections of lung cancer mortality in 4 industrialized countries. *International Journal of Cancer*, 117, 476–485. <https://doi.org/10.1002/ijc.21078>
- Sinha, R., Cross, A. J., Graubard, B. I., Leitzmann, M. F., & Schatzkin, A. (2009). Meat intake and mortality: A prospective study of over half a million people. *Archives of Internal Medicine*, 169, 562–571. <https://doi.org/10.1001/archinternmed.2009.6>
- Stoney, C. M., & Johnson, L. L. (2012). Design of clinical studies and trials. In J. I. Gallin & F. P. Ognibene (Eds.), *Principles and practice of clinical research* (pp. 225–242). Academic Press.
- Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology and Therapeutics*, 15, 443–453. <https://doi.org/10.1002/cpt.1974155443>
- Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children*, 71, 181–194. <https://doi.org/10.1177/001440290507100204>
- Thompson, R. S., Rivara, F. P., & Thompson, D. C. (1989). A case-control study of the effectiveness of bicycle safety helmets. *The New England Journal of Medicine*, 320, 1361–1367. <https://doi.org/10.1056/NEJM198905253202101>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Underwood, B. J. (1949). *Experimental psychology: An introduction*. East Norwalk, CT: Appleton-Century-Crofts.

- U.S. Department of Health and Human Services (2014). *The health consequences of smoking—50 years of progress: A report of the Surgeon General*. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health.
- U.S. Preventive Services Task Force. (1996). *Guide to clinical preventive services* (2nd ed.). Baltimore: Williams and Wilkins.
- Van der Geest, J. (2009a). LATSQ. (randomized) Latin Square. MATLAB script. <http://www.mathworks.com/matlabcentral/fileexchange/12315-latsq/content/latsq.m>
- Van der Geest, J. (2009b). BALLATSQ—Balanced Latin Square. MATLAB script. <https://nl.mathworks.com/matlabcentral/fileexchange/9996-ballatsq/content/ballatsq.m>
- Van Leeuwen, P. M., Happee, R., & De Winter, J. C. F. (2014). Vertical field of view restriction in driver training: A simulator-based evaluation. *Transportation Research Part F: Traffic Psychology and Behaviour*, 24, 169–182. <https://doi.org/10.1016/j.trf.2014.04.010>
- Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, 25, 195–217. <https://doi.org/10.1093/ijpor/eds021>
- Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101, 817–835. <https://doi.org/10.1037/a0016127>
- Wang, C., Lu, L., & Lu, J. (2015). Statistical analysis of bicyclists' injury severity at unsignalized intersections. *Traffic Injury Prevention*, 16, 507–512. <https://doi.org/10.1080/15389588.2014.969802>
- Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research*, 2, 149–168. <https://doi.org/10.1071/CH9490149>
- Wood, R. W. (1904). The n-rays. *Nature*, 70, 530–531. <https://doi.org/10.1038/070530a0>
- Zanotto, D., Rosati, G., Spagnol, S., Stegall, P., & Agrawal, S. K. (2013). Effects of complementary auditory feedback in robot-assisted lower extremity motor adaptation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 21, 775–786. <https://doi.org/10.1109/TNSRE.2013.2242902>

Human Subject Research for Engineers

A Practical Guide

de Winter, J.; Dodou, D.

2017, IX, 105 p. 23 illus., 9 illus. in color., Softcover

ISBN: 978-3-319-56963-5