Hypotheses and Models for Theory Testing 7

7.1 Hypothesis Testing and Significance Tests

Chapter 2 discusses the close connection between theory and hypothesis. A social science theory contains statements that can or should be verified empirically (Hunt 2010, pp. 188ff.). For such tests, **hypotheses** are central. As explained in Sect. 5.2, these are assumptions about facts (for example, the presence of certain characteristics) or relationships (for example, between attitude and behavior). Now, such general assumptions often cannot be tested at a general level. For example, in the case of the relationship between attitude and behavior, it is often necessary to define specific persons, countries, dates, etc. for an empirical investigation. Hence, one deduces more concrete hypotheses (related to specific situations) from the general theoretical statements (→ deduction, see Sect. 2.5). This is the first step of operationalization, which is followed by the development of measurement instruments, selection of study participants, data collection and data analysis.

In general, one can say that scientific hypotheses are assumptions about facts that go **beyond** the individual case and that can be empirically tested. They represent a link between theory and empiricism. According to Bortz and Döring (2006), the requirements for hypotheses are:

- **Empirical examination**: Scientific hypotheses must relate to real facts that can be empirically investigated.
- **Conditionality**: Scientific hypotheses must be based, at least implicitly, on a meaningful “if-then-proposition” or a “the-more-the-more-proposition”. In this sense, assumptions about facts are implicit conditionals. For example, the assumption that, “at least 10% of under-30 year olds have not completed vocational training” can be formulated as, “if a person is under the age of 30, the probability of not having completed vocational training is at least 10%”.

M. Eisend, A. Kuss, *Research Methodology in Marketing*,
https://doi.org/10.1007/978-3-030-10794-9_7

- **Generalizability**: Scientific hypotheses must make statements beyond the individual case or a singular event.
- **Falsifiability**: Scientific hypotheses must be (empirically) refutable.

Here are three examples of hypotheses in the field of marketing research:

"The ease of use of a piece of software improves the satisfaction with the software by its users."

"The relationship between attitude and behavior of consumers becomes stronger when the behavior is socially desirable."

"The longer a business relationship lasts, the lower the likelihood that a partner will end the relationship in the near future."

In the context of hypothesis testing using statistical methods, it is important to understand the distinction between alternative and null hypotheses. The **alternative hypothesis** is the statistical formalization of the research question. It is formulated as a statistical assumption that there will be effects, differences, or relationships. The **null hypothesis** contradicts the alternative hypothesis. Research studies usually try to confirm effects (alternative hypothesis). Therefore, the null hypothesis assumes that there are no effects and a hypothesis test attempts to reject the null hypothesis. If it is rejected, one decides to accept the alternative hypothesis. If the null hypothesis cannot be rejected, it will be retained. An alternative hypothesis such as, "satisfied customers are more likely to recommend a product" would be formulated as a null hypothesis thus: "satisfied customers are not more likely to recommend a product". If the hypothesis test shows that this null hypothesis cannot be rejected, then it follows that there is no relationship between customer satisfaction and the likelihood of the customer recommending a product.

A hypothesis is supported if its statement and the corresponding empirical observations are in agreement. However, what does "agreement" mean? The problem of such decisions is illustrated by the following examples:

- We assume (hypothesize) that after at least 10 contacts with brand messages, consumers will actively remember that brand. A study with 200 subjects shows that this was the case for 160 people, but not for the remaining 40 people. Is this result consistent with the assumption?
- We assume (hypothesize) that the intensity of the post-purchase service determines customer satisfaction. In a related study, there is a correlation between these two variables of $r = 0.42$, well below $r = 1.0$ (i.e., a perfect correlation). Is the hypothesis supported?
- We assume (hypothesize) that there is no correlation between the variables "age" and "interest in ecological products", i.e., that the corresponding correlation is at $r = 0$. However, when we investigate the relationship, we find a correlation of $r = 0.08$. Is there a relationship between the two variables?

The questions raised in the first two examples can be clarified easily based on the considerations concerning scientific explanations (see Sect. 2.3.2). Obviously, the first example is not concerned with a regularity which applies to each individual case (→ deductive nomological explanation), but with a *statistical-relevance explanation* that refers to a probability statement (in this case with regard to brand memory). In the second example, we cannot assume that only one variable (post-purchase service) influences another variable (customer satisfaction). Since only one out of a larger number of influencing factors is considered, the relationship between both variables is not perfect or deterministic. Therefore, the resulting correlation is clearly less than 1.0. Rather, in the sense of an explanation based on statistical relevance, we empirically examine whether a substantial correlation (correlation distinctly different from 0) exists between the variables, which would probably be confirmed in the example.

Now to the third and somewhat more complicated example. Here, the question of "**significance**" becomes particularly obvious, that is, the question of whether there is a systematic difference between the expected correlation (r = 0) and the measured correlation (r = 0.08). The significance or significance level indicates the probability that, in the context of a hypothesis test, the null hypothesis ("there is no systematic relationship") can be erroneously rejected, even though it is actually correct (Type I error, see Sect. 7.3). Therefore, the level of significance is also referred to as the **error probability**. In order to answer the question of significance, we apply *inferential statistics* that serve to make decisions on such questions. In the example case, if one were to take into account the difference between the two values—the desired confidence interval and the sample size with respective distribution assumptions—such a decision could be made. The p-value commonly used for such decisions indicates in this example how large the probability is that a value r = 0.08 will be found in the respective sample, if in the population the (actual) value is r = 0 (Sawyer and Peter 1983, p. 123). It becomes clear that this is an inductive reasoning, from a relatively small number of cases to an often very large population (for example, the entire population of a country).

A schematic application of statistical methods only for hypothesis tests would be too simple, because all possible errors due to operationalization and measurement would be completely ignored. Such **systematic errors** can be much more serious than sampling errors.

From the point of view of scientific realism (see Chap. 3), one has yet to draw attention to another problem in significance tests. These tests summarize group differences or relationships between variables in a single measure. For example, we may find a positive relationship between variables A and B in 70 or 80% of the subjects studied, but for the remaining individuals, this relationship may be absent or may even be a negative one. However, one would interpret a significantly positive correlation coefficient as having confirmed a suspected positive association. A summary review of several such results would reinforce this effect, giving the impression that these results are quite homogeneous and unambiguous. From the perspective of scientific realism, however, it would make sense to contrast the "empirical successes" with the "empirical failures" (see Sect. 5.3). This aspect is an argument for conducting meta-analyses (see Sect. 9.3).

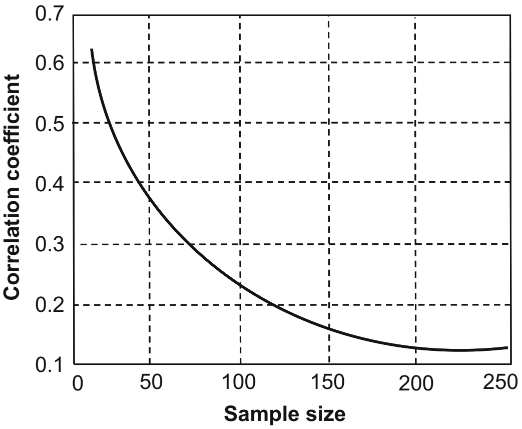
There are also associations between variables that have no logical relationship, so-called **spurious correlations**. A popular example is the empirically observable relationship between the number of storks and the birth rate in different regions. The reason for this association is obviously a third variable: In the countryside, where there are more storks, there are also more families with many children living there.

> W. Lawrence Neuman (2011, p. 413) gives the following assessment of the importance of significance tests:
>
> "Statistical significance tells us only what is likely. It cannot prove anything with absolute certainty. It states that particular outcomes are more or less probable. Statistical significance is not the same as practical, substantive, or theoretical significance. Results can be statistically significant but theoretically meaningless or trivial. For example, two variables can have a statistically significant association due to coincidence with no logical connection between them (e.g., length of fingernails and ability to speak French)."

7.2 Statistical Versus Substantial Significance

The problem discussed above leads to the crucial comparison between **statistical significance** and **substantial significance**. Whether or not a statistical significance occurs depends on various influencing factors. A central influencing factor is the number of cases of an investigation. Figure 7.1 illustrates this relationship using the example of the correlation coefficient. The larger the sample size, the smaller the correlation coefficient, which satisfies the significance criterion of $p < 0.05$, which is frequently used as a critical threshold in marketing research.

Fig. 7.1 Relationship between sample size and correlation coefficients being significantly different from zero at $p < 0.05$



For very large samples, highly significant results can be found, but they may have only minor theoretical or practical relevance, since the size of the considered effect is very small. In fact, if a sample is large enough, almost every result (e.g., a correlation coefficient) that is only slightly different from zero (or any other comparison value) would be significant. Statistical significance is thus a *necessary but insufficient criterion for a practically or scientifically relevant statement* (that is, for **substantial significance**). For the assessment of the relevance of the hypothesis, effect size is an important criterion that does not dependent on the sample size. We already addressed this problem in Sect. 2.3.2, where an example is provided of the significant difference between substantial and statistical significance.

If statistical significance is given, then the substantial significance can be assessed by the effect size. An **effect size** is "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (Kelley and Preacher 2012, p. 140). According to Kelley and Preacher (2012), effect sizes can have different dimensions (e.g., variability, association) and these dimensions are operationalized by different effect size measures or effect size indices. For instance, the dimension of variability can be operationalized in units of variance or standard deviations, the dimension of association in units of correlations. When an effect size measure is applied to data, we obtain an effect size value (e.g., a correlation of 0.25). Marketing research often applies effect sizes that express **associations between variables** or the **strength of relationships** (such as correlations) that indicate how strongly two variables are related, and how large the explanation of the variance of a dependent variable by an independent variable is (Eisend 2015). These effect sizes are commonly applied to describe the relationship between two variables, although some measures of *explained variance* exist that describe the size of an effect that can relate to more than one independent variable and one dependent variable. Effect sizes that measure the extent of the variance explained are central to science, which is above all concerned with explanations. The more science can explain, the better (Aguinis et al. 2011).

The following simple example illustrates the importance of substantial significance and the questionable use of statistical significance in large samples:

A correlation coefficient of 0.03 that measures the relationship between income and happiness is significant at $p < 0.05$ in a sample of 10,000 participants. The result indicates that income is significantly related to happiness (statistical significance). However, the correlation coefficient corresponds to a proportion of explained variance of ca. 0.1 percent. That means 99.9 % of the variation in happiness or income (whatever we apply as dependent variable) remains unexplained, which would be a disappointing figure for scientists who want to explain differences in income or happiness (i.e., its substantial significance).

While significance testing attempts to answer the question of whether there is any difference, effect, or correlation between two variables, the effect size dimension that refers to the strength of a relationship indicates how close the relationship is between two variables. The effect size can be used not only to describe the relationship between two continuous variables (e.g., income and happiness), but also to describe the relationship between two binary variables (e.g., gender and whether someone is a smoker). Although statistical tests for such variables focus on finding out differences or separation (e.g., whether there are more male or female smokers), the test can be understood as one that describes the relationship between gender and smoking. Thus, effect sizes that describe relationships between two variables are appropriate. The effect size shows more meaningful results than a significance test for various reasons:

- Effect sizes *can be compared across different studies and across different types of variables*. This is a common approach in medical science when comparing different studies that examine the effect of different procedures (e.g., type of medication, hospital treatment time, and alternative therapy) in curing the same disease. The higher the explained variance due to a particular procedure (i.e., the closer the association between a procedure and the curing of the disease), the more successful the procedure is.
- Effect size measures such as correlations are often easy to interpret and therefore *more comprehensible to practitioners* than significance tests.
- Effect sizes provide meaningful "*benchmarks*" for comparisons with other study results, between disciplines or even between researchers (Eisend 2015).
- Finally, for each effect size, confidence intervals can also be reported that provide an equivalent to significance tests: if the confidence interval does not contain zero, then the effect is significant (Lipsey and Wilson 2001).

The trend towards "big data" in research, i.e., increasing amounts of data, mainly due to the use of digital technologies, and the issue of the problems of statistical testing on very large samples, has already been addressed in Sect. 6.4. This also explains the increasing importance of effect sizes compared to significance tests. Therefore, a number of scientific journals are placing increasing emphasis on reporting effect sizes while devaluing the importance of significance tests, for example, the "*Strategic Management Journal*" (see Bettis et al. 2016). The journal *Basic and Applied Social Psychology* has even decided not to allow any significance tests (Trafimow and Marks 2015). The current use of significance tests also encourages researchers to engage in dubious practices (e.g., p-hacking, see Chap. 10) to reach results that meet the required significance levels, thus increasing capitalization on chance, biasing the scientific knowledge base and diminishing the probability that results are reproducible (Aguinis et al. 2017).

Another important effect size in marketing research is the **magnitude of an effect**, which provides important information from a substantive and applied perspective. In contrast to the effect size dimension referring to the strength of a relationship, the magnitude of an effect usually applies effect size measures for the *relative change of a dependent variable Y with respect to a relative change of an independent variable X* (e.g., elasticity). This effect size is often of great practical relevance in marketing research because it provides information for an input-output analysis. For example, it can be applied to determine the relative increase in sales of a product due to the relative increase in advertising spending.

The two most important effect size dimensions that are used in marketing research (strength of the relationship and the magnitude of the effect) can thus be distinguished as follows:

- The **strength of a relationship** indicates *how close* the relationship between variables is. Common measures are—amongst others—correlations or proportion of explained variance (see above).
- The **magnitude of an effect**, on the other hand, represents the *extent of change* in a dependent variable due to the change of an independent variable. Common indicators are (unstandardized) regression coefficients.

Both aspects should be additionally illustrated by the example of a linear regression with one dependent and one independent variable (see Fig. 7.2). It shows some (fictitious) measurements and a corresponding regression line. The slope of this line is indicated by a triangle. This slope indicates the magnitude of the effect. Furthermore, the distances of the actually observed values of the dependent variable (y) from the values expected based on the respective x-values and the regression

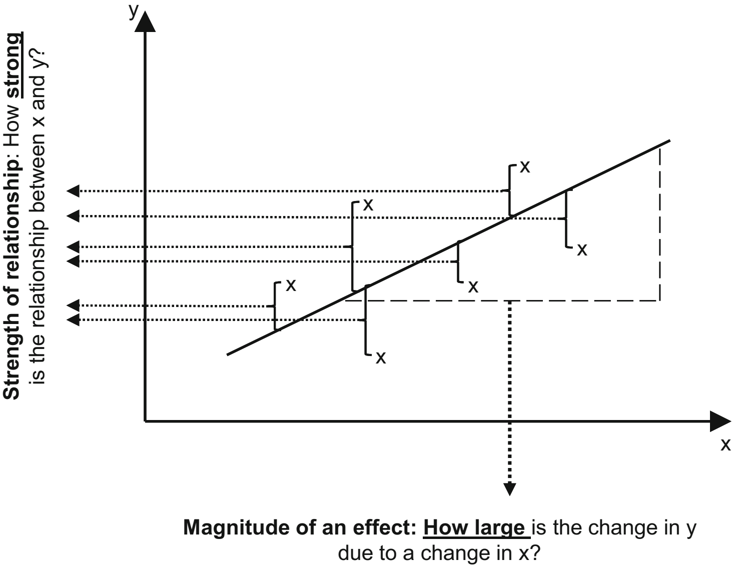
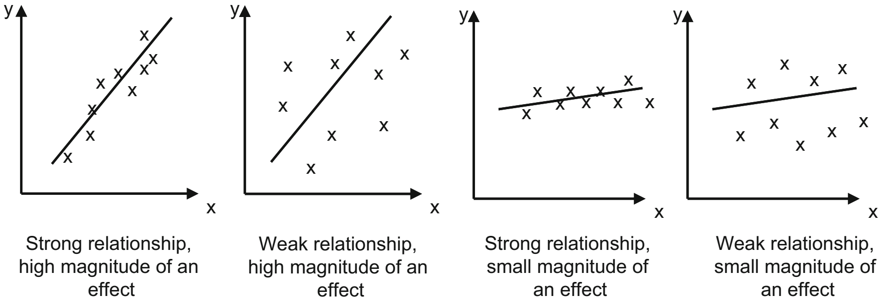

Strength of relationship: How strong is the relationship between x and y?
y
x
x
x
x
x
x
x
x
Magnitude of an effect: How large is the change in y due to a change in x?

Fig. 7.2 Strength of relationship versus magnitude of an effect



Strong relationship, high magnitude of an effect

Weak relationship, high magnitude of an effect

Strong relationship, small magnitude of an effect

Weak relationship, small magnitude of an effect

Fig. 7.3 (No) equivalence of strength of relationship and magnitude of an effect

relationship are entered. The smaller these distances are, the stronger the relationship between the variables x and y seems to be.

The strength of a relationship and the magnitude of an effect are therefore not equivalent: a large effect can occur even with little explained variance and thus a weak relationship between two variables. In Fig. 7.3, the difference is illustrated by the relationship between two variables.

7.3 Power of Statistical Tests

The relationship between significance tests, sample sizes and effect sizes is taken into account in the context of "**power analysis**" (Cohen 1988). This analysis addresses the problem of making two mistakes in testing hypotheses:

- One erroneously rejects the null hypothesis. This is a **Type I error**, that is, the mistake of rejecting the null hypothesis, even though it is actually correct. The probability of this is determined by the level of significance or the error probability.
- One erroneously assumes the null hypothesis. This is a **Type II error**, the mistake of accepting the null hypothesis, even though it is actually wrong.

The four possible results of a significance test are depicted in Fig. 7.4.

The smaller the α-error in a study, the less frequently the null hypothesis is falsely rejected. This increases the probability of mistakenly accepting the null hypothesis

	H_0 is true	**H_0 is false**
H_0 not rejected	Correct decision	Type II error (β-error)
H_0 rejected	Type I error (α-error)	Correct decision

Fig. 7.4 Results and errors of hypothesis testing

and rejecting the alternative hypothesis (β-error). However, the size of the α-error does not directly deduce the size of the β-error and vice versa. The two types of errors are determined in different ways. The size of the α-error depends on the significance level.

The size (1-β) is also referred as **power** (Cohen 1988). The power of a test (that is, the likelihood that testing a null hypothesis leads to rejection of the null hypothesis if the alternative hypothesis is correct) is influenced by three factors (next to the variance):

- α-significance level: the smaller α, the lower the probability of choosing the alternative hypothesis falsely (Type I error);
- Sample size: the larger the sample size, the greater the probability of deciding in favor of the alternative hypothesis (ceteris paribus);
- Effect size: the larger the explained variance and the strength of a relationship, the greater the power of the test and thus the probability of deciding against the null hypothesis and in favor of the alternative hypothesis.

In summary, at a given significance level (e.g., $\alpha = 0.05$) larger effect sizes tend to become more likely significant than smaller effect sizes and larger samples have higher test sensitivity than small samples, and thus are more likely to produce significant results.

Although there are no formal **standards for power levels** (also referred to as π (pi)), a value of $\pi = 0.80$ is usually used, that is, a four-to-one probability between β-error and α-Error (Ellis 2010). If the test is designed in such a way that it should not produce any β errors, then a lower standard can be applied. This is often the case in medical research, where it is better to assume that one has an indication of a disease, even if the patient is healthy, than to assume that a patient is healthy, but in reality, is suffering from a disease.

Power analysis is important for the *interpretation of test results*, because the power indicates the probability of correctly rejecting the null hypothesis. It is, as already explained, dependent on the chosen significance level, the effect size, and the sample size. This attests to the central idea that a hypothesis can be rejected for various reasons. A hypothesis may be rejected because the effect is too small, which is easy to understand and desirable from a scientific point of view. However, a hypothesis can also be rejected because the sample is not large enough or the significance level is too small, that is, it was chosen as being too strict. With an increase in the sample size or a "more generous" level of significance, the hypothesis could possibly be accepted based on the same data.

How to choose the *right significance level*? The social sciences have established a significance level of 5%, proposed by Ronald Fisher (1925, p. 43). This limit means that, on average, one in twenty studies in which the null hypothesis is correct (e.g., age is not related to happiness) is found to be false (e.g., age is related to happiness). Sometimes results are accepted even at a lower significance level of <0.1. Which levels of significance are accepted also depends on the degree of innovation of a study: scientists tend to apply less stringent criteria to completely new and innovative results, and possibly consider marginally significant results to be relevant, than to results that relate to an already established hypothesis. Depending on the object under investigation, a Type I error may be less serious than a Type II error, as indicated in the above example in medical science, where one is more likely to accept a disease, even if the patient is healthy, than to assume that a patient is healthy, when she or he is actually ill.

The relationship between level of significance, effect size, and sample size also makes it possible to *determine the sample size* for a known or expected effect size that is necessary so that the effect at a given level of significance with a desired power is actually significant. It can already be seen in Fig. 7.1 that large effect sizes require smaller samples in order to reach the specified significance level and vice versa. In addition, if the power level is high, the sample size needed to reach the significance level continues to increase, especially with small effect sizes.

7.4 A Priori Hypotheses Versus Post Hoc "Hypotheses"

The usual applications of statistical tests are based on a procedure in which one hypothesis or a few specific hypotheses are formulated, then appropriate data are collected and suitable statistical tests are applied. Examples include studies on the efficacy of drugs (new drug vs. placebos) or testing previously theoretically well-founded hypotheses (such as the relationship between x and y). Figure 7.5 illustrates this "classical" approach to testing (a few) **hypotheses that have been formed a priori**. It shows the path from a few theoretically well-founded hypotheses to data

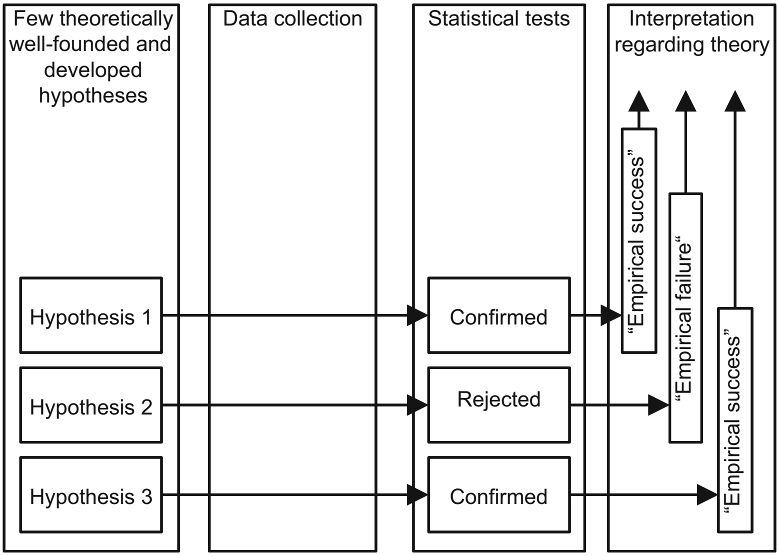


Fig. 7.5 Procedure of testing a priori hypothesis

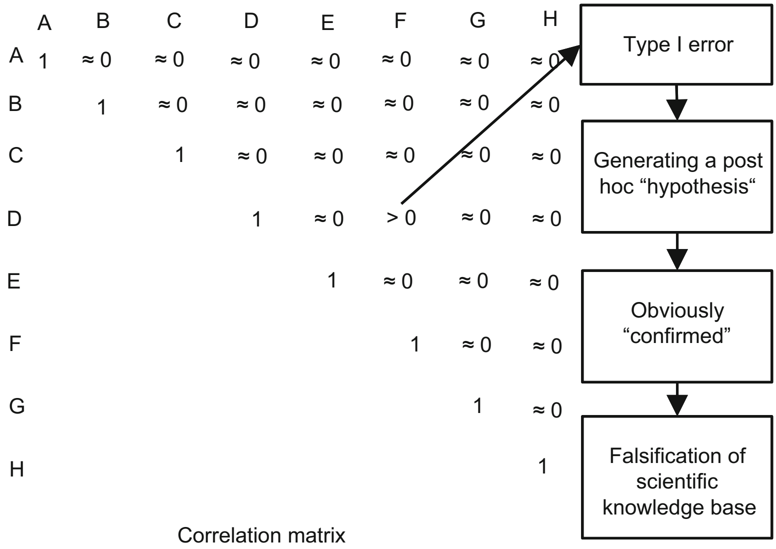
collection and statistical tests that confirm or reject the hypotheses and to their interpretation as "empirical successes and failures" (see Sect. 5.3). This procedure also corresponds to the hypothetico-deductive method described in Sect. 5.2.

In many marketing research studies, however, data collection is not limited to very few selected variables. It is more common to include a larger number of variables. For example, survey questionnaires usually include a two-digit number of questions with a corresponding number of variables. Under these conditions, researchers may choose from a variety of possible (and easy to compute) correlations those that appear to be "significant", and hypothesize the relationships later, because it is easier to publish significant results than non-significant ones (see Sect. 10.2.4). With the goal of increased publication opportunities (which is, under today's publication pressure, partially understandable), theories and hypotheses are adapted to already existing results; that is, **post hoc hypotheses** are formulated. These are not real hypotheses (see Sect. 7.1) because, given already existing results, one cannot speak of assumptions and falsifiability is not possible. The problem is not unknown in the literature (and probably also in research practice): Peter (1991, p. 544) speaks of "fishing through a correlation matrix"; Kerr (1988) speaks of "**HARK**ing: **H**ypothesizing **A**fter the **R**esults are **K**nown"; Leung (2011) discusses "Presenting Post Hoc Hypotheses as A Priori ...". Already some fifty years ago, Selvin and Stuart (1966) referred to such an approach as "data dredging". The extent of the problem in research practice is difficult to know, because in such cases, the authors avoid disclosure and readers of articles based on HARKing find few clues. Banks et al. (2016) reported in a study that about 50% of respondents to a survey in management research said they had "presented a post hoc hypothesis as if it were developed a priori" (p. 10). The problem concerns research ethics (see Sect. 10.2.4) and can lead to grossly misleading results. The reasons are briefly outlined below.

The starting points of the considerations are the following real-life experiences:

- Researchers are anxious to find significant results because their chances of publication are much greater than for those of non-significant ones.
- For a larger number of potential associations of variables, by chance, some seemingly "significant" relationships arise, even if no such relationships actually exist. Even if one correlates numerous variables, which were generated by random numbers, for which there can be no systematic relationship, a few correlation coefficients would be "significantly" different from zero and misleadingly indicate that there are real relationships (Kruskal 1968).

The problem is illustrated by a very simple example. Figure 7.6 shows a (hypothetical) correlation matrix for the variables A to H, which are measured in a reasonably large sample. In the corresponding population, there is no correlation between any of these variables, so that the corresponding correlation coefficients are (or should be) 0. Accordingly, in the correlation matrix for the (sample) data, in the main diagonal are the "1" values and in the other fields are values which are very close to 0 (ideally the value 0). However, it may well be that through sampling, some cases were sampled that led by chance to some correlation coefficients that are



	A	B	C	D	E	F	G	H
A	1	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
B		1	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
C			1	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
D				1	≈ 0	> 0	≈ 0	≈ 0
E					1	≈ 0	≈ 0	≈ 0
F						1	≈ 0	≈ 0
G							1	≈ 0
H								1

Correlation matrix
Type I error
Generating a post hoc “hypothesis“
Obviously “confirmed”
Falsification of scientific knowledge base

Fig. 7.6 Procedure of generating post hoc “hypotheses”

clearly greater than 0 (thus apparently “significant”). In the example this is entered for the variable combination D and F, which is marked with “> 0”. This would correspond to a Type I error (see above), because the correct null hypothesis would be rejected. If, following from this result, a (post hoc) “hypothesis” is proposed, then its (apparent) confirmation would be unavoidable because the corresponding result is already known. Kerr (1988, p. 205) uses the ironic phrase “HARKing can translate type I errors into theory”. Furthermore, for hypotheses that have been formed subsequently, the requirement (see Sect. 7.1) that hypotheses can be rejected by the investigation is violated. The interpretation of such a random result as a statistical confirmation of a previously theoretically developed hypothesis would be misleading in regard to the relevant scientific knowledge.

An example of a problematic use of significance testing is a study concerning personality traits and consumer behavior that appeared in the early years of the highly respected *Journal of Marketing Research*. The study was about relationships between personality traits (e.g., aggression) and consumer behavior. For this purpose, the relationships between three personality variables and 15 characteristics of consumer behavior (product use, brand preferences in different product groups) were examined (with a relatively weak database) by means of Chi^2 tests. In these 45 tests, there were seven (apparently) significant relationships. For example, an association has

(continued)

emerged between aggressiveness and the preference for wet or electric shaving, a connection that may not be theoretically compelling. It is questionable which proportion of the seven "significant" results has a real basis or came about by chance.

To distinguish from such an approach is the test of so-called **implicit hypotheses**. These are hypotheses that do not belong to the core of the theoretical question and are not necessarily fixed a priori (e.g., fixed in writing). However, for these hypotheses, the researcher collects corresponding additional data due to his or her experience and theoretical training, which suggests that there might still be interesting or relevant relationships (e.g., as a control variable). This would lead to a rather small number of additional hypotheses for which the statistical problem outlined above appears only to a limited extent. One may well assume that the "temptation" to HARKing is greatest when large (many variables) data sets, that are not self-collected, are used. On the other hand, in the case of one's own data collection, one usually deals with a restricted number of variables that were considered meaningful and important at the *beginning* of the investigation and then collected. The least likely is the problem in experimental studies (see Sect. 8.3), which is confined to a small number of carefully established variables.

It goes without saying that the description and documentation of particularly interesting results, which are not based on previously developed hypotheses, are of course possible, but not with the claim of statistical confirmation. If post hoc hypotheses are to be verified empirically / statistically, then another data set is required that is independent from the data from which this hypothesis was created. Furthermore, the interpretation of data without a priori hypotheses can make sense when applying an *inductive approach*. In any case, researchers need to be transparent about what they do. The problem of HARKing mainly refers to a lack of transparency, that is, when researchers present post hoc hypotheses as a priori hypotheses without acknowledging having done so (Kerr 1988).

7.5 Modeling with Regression Analysis

In the context of testing theories, some areas of marketing research use mathematical models for the presentation and solution of problems. Mathematically formalized modeling is also simply referred to as **modeling**. This approach is based above all on econometrics, a branch of economics that combines economic theory, empirical data and statistical methods. The central task of econometrics is the derivation of econometric models from economic theories and their numerical concretization. With the help of econometrics and modeling, interesting relationships in economics can be quantified (e.g., percentage change of the savings rate with percentage change of the interest rate), thus hypotheses and whole models can be empirically tested and these empirically validated models can be used for *forecasts* or *simulation* (e.g., economic growth will change as inflation rates change).

In marketing research, in addition to the applications of econometrics, *optimization questions* are often at the forefront of modeling. Shugan (2002) distinguishes between two different definitions of mathematical models, one being the mathematical optimization of variables and the other mathematical mapping with the purpose of solving research questions. In the former view, it is often sufficient to show that a particular solution is optimal, for example, what is the optimal ratio between advertising spending and personal selling? It is often about optimizing resource allocations. In addition to such models, which are oriented towards solving practical problems, they also serve to develop a theoretical understanding of marketing problems by varying assumptions and determining the resulting changes in dependent variables. Often the second approach does not involve a systematic empirical review of the model assumptions, but a fair presentation of the adequacy and successful application of such models based on selected cases is very common.

Parameterization and validation, in the context of modeling, use methods that are based on classical **regression analysis**. Regression analysis is a statistical method that attempts to explain the change in a *dependent variable* by changes in a set of so-called *explanatory* or *independent variables* by quantifying a single equation. A regression can determine whether there is a quantitative relationship between the independent variables and the dependent variable. However, the result of a regression analysis alone cannot show causality even when statistical significance is given, since a statistical relationship never implies causality (for causality and the special requirements for the appropriate study design, see Chap. 8). Nevertheless, regression analysis and other econometric techniques are used to determine relationships between variables, which are often interpreted as cause-and-effect relationships. In order for the empirical regression analysis to be done, strict assumptions must be fulfilled.

In the simplest case, a regression model $Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon$ describes an endogenous variable Y by a linear relationship to one or more other (exogenous) variables $X_1, ..., X_k$. The model explains the **endogenous** variable, while an **exogenous** variable is explained, not by the model, but by variables outside the model. For instance, if the regression model tries to explain the degree of happiness by variables such as age and income, happiness is an endogenous variable in the model, while age and income are exogenous variables. Of course, income can be explained by other variables such as education, but since they are not included in the model, income is considered an exogenous variable in that particular model. Since in practice there will be no exact relationship between the empirically observed variables, so that the exogenous variables could fully explain the endogenous variable, an error term in the equation records all variables that besides $X_1, ..., X_k$ also have an influence on Y. After specifying a particular model, the model parameters $\beta_0, ..., \beta_k$ are estimated. On this basis, forecasts can be made for the values of Y for assumed values of $X_1, ..., X_k$.

Regarding the usual results of a regression analysis, we find results that stand for significance and measures for effect sizes in terms of the strength of a relationship and magnitude of the effect (see Sect. 7.2):

- **Strength of a relationship/explained variance**: The corresponding measure R^2 *(coefficient of determination)* shows what proportion of the variance of the dependent variable is explained by all the independent variables.
- **Magnitude of an effect**: The *unstandardized regression coefficients* β_0, ..., β_k indicate how much a change of the respective independent variable affects the dependent variable, that is, by what extent the dependent variable changes, if the independent variable changes by a certain extent. This value depends on the scaling of the variable. Thus, for example, the magnitude of the effect that measures the relationship between advertising spending and sales (units sold) depends on whether we measure the spending in US dollars, euros, or Swiss francs. If these coefficients are specified as elasticities, that is, the ratio of the percentage change in one variable (e.g., sales) to the percentage change in another variable (e.g., advertising spending), the scaling problem is eliminated.
- **Significance of the regression model:** Tests are used to check whether the proportion of explained variance (R^2) is significantly different from 0, that is, whether the model makes (at least a small) contribution to the explanation of the dependent variable (see also Sect. 2.3.2).
- **Significance of the regression coefficients**: With t-tests, we check whether the different regression coefficients β are significantly different from 0. Otherwise—at $\beta = 0$—a change in the respective independent variable would have no systematic effect on the dependent variable.

The standard method for estimating the parameters in linear regression models is the **Ordinary Least Squares (OLS)** estimation. In order to be able to apply it without problems, a number of assumptions have to be fulfilled, which also have important substantive implications with regard to theory testing (Allison 1999; Hair et al. 2010):

- The regression model must have *parametric linearity* (i.e., the relationship of the variable must follow a linear function) and not all observations of an X variable may be the same (i.e., they must vary), otherwise no estimation is possible.
- The conditional expected value of the error term must be zero, which implies a covariance between the X variables and the error term of zero. This assumption of the *exogeneity* of X_1, ..., X_k is important, because only in this case are ceteris-paribus statements, such as “a change of X_1 by one unit leads to a change of Y by β_1 units,” possible. For instance, the influence of advertising spending on sales can lead to endogeneity problems, because advertising spending decisions often depend on sales in prior periods and are therefore not exogenous to the model. A statement such as “a change of 10% in advertising spending leads to a change of 3% in sales” would be wrong, since the change in sales also depends on the sales of the prior period, as does the change in advertising spending.

- The conditional variance of the error term must be constant. A famous example for a violation of this condition is the relationship between income and spending on certain consumption activities, such as food. At a low income, consumers spend a certain constant amount on food, as they cannot afford more. With increasing income, consumers display a greater variation in spending on food, as they sometimes buy inexpensive food but at other times enjoy expensive meals. As a result, the error term variance would increase with the increase in the independent variable.
- The conditional error term covariance must be equal to zero, which means that the data point deviation from the regression line does not show any pattern along the independent variable. This is often violated in *time series data*, where the independent variable is time. Most data points show a particular pattern over time, for example an economic cycle, and a data point is not independent of the preceding data point (e.g., if the economy shows high economic growth in one year, it probably shows relatively high economic growth in the following year, too). As a result, the error terms show a co-variation pattern.
- There must be no perfect correlation between the explanatory variables, since in this so-called perfect *multicollinearity* an OLS estimation is impossible. In addition, imperfect multicollinearity, characterized by high correlations between explanatory variables, is problematic, because in this case OLS cannot precisely distinguish between the influences of the individual variables and cannot provide accurate parameter estimates.
- The error terms should be normally distributed.

One can use a number of statistical tests to obtain evidence for a *violation of these assumptions*. When violations are identified, the model specification can be revised, robust procedures can be used, or alternative estimation techniques (such as instrumental variables) can be used, depending on the nature of the problem. If the theory already suggests that assumptions of the classical regression model are not realistic (e.g., a correlation of the error terms occurs regularly with time series data), alternative estimation methods are usually used right from the start. The following is a brief illustration of how to deal with the violation of the respective assumptions (for more detail, see Allison 1999 or Gujarati 2003).

- If the assumption of the **parameter linearity** is not met, a parameter-linear form can be produced by variable or model transformation (for example by log transformation). Meanwhile, there are also estimation methods for non-linear relationships (non-linear least squares).
- **Endogeneity** can be detected with the Hausman test. To solve the endogeneity problem, one can introduce an instrumental variable (IV estimation). This requires so-called instrumental variables that are highly correlated with the endogenous explanatory variables (instrument relevance) and at the same time are not correlated with the error term (instrument exogeneity). Given the proper quality of the IV estimator, consistent parameter estimates are achieved. The

quality of the instruments can be checked by regressing the endogenous explanatory variable on all instruments, including the exogenous variables.

- Whether or not the problem of **heteroscedasticity** occurs (i.e., not a constant conditional variance of the error term) can also be tested, using either the Breusch-Pagan or White test. In the case of heteroscedasticity, robust error terms can be used instead of the standard error terms, which the OLS wrongly estimates. Alternatively, the use of WLS (Weighted Least Squares) is conceivable in large samples.
- In time series regressions (i.e., data are collected repeatedly at different points in time), one often faces the problem of error term **autocorrelation**, which is detected by various tests (Durbin-Watson test, Breusch-Godfrey test). Again, one has the opportunity to use autocorrelation robust standard errors or to estimate a GLS (Generalized Least Squares) model. This procedure provides correct standard errors, and more efficient estimates of the model parameters, if the autocorrelation structure used for the model transformation is correctly recognized and implemented in the new model.
- Perfect **multicollinearity** is unlikely to occur in social science research, but high multicollinearity can occur. High multicollinearity is often recognized by high pairwise correlations between the independent variables and high coefficients of determination in models, in which one exogenous variable is explained by all other exogenous variables. The Variance Inflation Factor (VIF), or Tolerance, measures multicollinearity. High multicollinearity is avoided by excluding variables from the regression model or by grouping variables into factors or indices.
- The **assumption of the normal distribution of the error term** is usually not subject to intensive tests in practice. Due to sufficiently large samples, a normal distribution of the estimated parameters can be assumed due to the central limit theorem.

7.6 Structural Equation Models

Structural equation models work as a test of networks of hypotheses or larger parts of theories. The alternative designation of **causal models** is somewhat problematic, similar to regression analysis, because the application is often based on cross-sectional data that do not allow a proof of causality. This is outlined in Chap. 8: "The ability to make a causal inference between two variables is a function of one's research design, not the statistical technique used to analyze the data that are yielded by that research design." (Jaccard and Becker 2002, p. 248). Finally yet importantly, it is difficult to exclude alternative explanations for a common variation of causes and effects (see Sect. 8.1).

The basic idea of Structural Equation Models (SEM) is that, based on the variances and covariances of indicators (observable variables) found in a dataset, conclusions are drawn with respect to *relationships between complex constructs* (latent variables). The characteristic features of structural equation models can be seen in the fact that a larger number of interconnected relationships is analyzed, and at the same time not directly observed concepts could be included in these relationships, whereby measurement errors can be explicitly taken into account.

The following is an illustration of the simultaneous analysis of multiple relationships, whereby possible measurement errors are not taken into account. The underlying model is the Technology Acceptance Model (TAM) of Davis et al. (1989), widely applied in technology use research, which explains the acceptance and use of (computer-based) technologies. A simplified model is depicted in Fig. 7.7. It assumes that the intention to use a technology depends on the perception of the usefulness of this technology (H1) and the ease of use (H2). The ease of use also influences the perceived usefulness (H3). Intention to use increases the actual use (H4). It can be seen that in this model several hypotheses or a part of a theory are *simultaneously* considered and (later) tested.

Such a model is called a structural model. It describes *relationships between the latent variables* (concepts). These variables cannot be observed directly, but can be estimated using appropriate measurement models. The next step is the development and application of these measurement models (similar to scale development, see Sect. 6.2), so that the parameters of the model can be estimated. For this purpose,

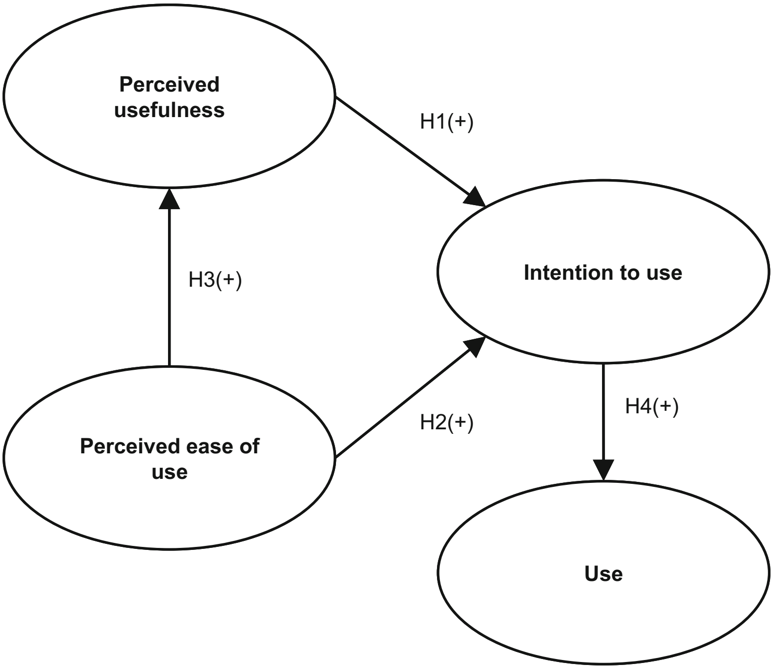


Fig. 7.7 Example of a structural model (simplified Technology Acceptance Model by Davis et al. 1989)

different **indicators** are used in the present example for the different latent variables. For example, the perceived usefulness of a technology can be measured with the following indicators. Respondents indicate the extent to which they agree with these statements on a scale ranging from 1 (“totally disagree”) to 7 (“fully agree”) for the endpoints:

- Productivity: “Using this technology facilitates productivity.”
- Effectiveness: “Using this technology facilitates effectiveness.”
- Performance: “Using this technology improves performance.”

Accordingly, all latent variables are measured by appropriate indicators (all are *manifest variables*). The (simplified) representation of the structural model with the corresponding measurement models is depicted in Fig. 7.8.

Measurement errors are considered in such models in two ways: Each indicator (e.g., “productivity” or “effectiveness”) is associated with a measurement error that is unobservable. The idea behind it is analogous to a regression model. In that, the latent variable explains the indicator, with the measurement error added like an error term in the regression analysis, because the explanation is not complete. Similarly, endogenous latent constructs (that is, variables that are explained by other constructs in the model, e.g., “intention to use”) are each assigned a measurement error that captures the unexplained variance next to the explained variance by the constructs influencing them (e.g., “perceived ease of use”).

The measurement error design is due to indicators that are called **reflective indicators**, that is, indicators that are caused by the latent variable. Accordingly, the arrows in the model are directed so that the latent variable and the measurement error explain an indicator. Thus, it is assumed that the latent variable (e.g., “perceived usefulness”) causes the different expressions of the indicators (“productivity”,

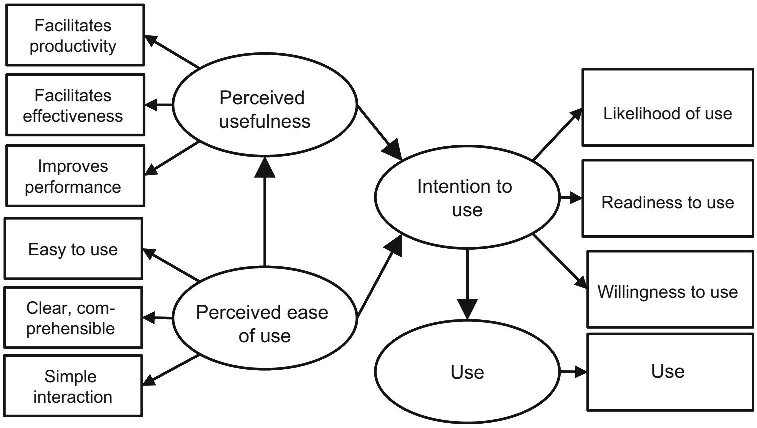
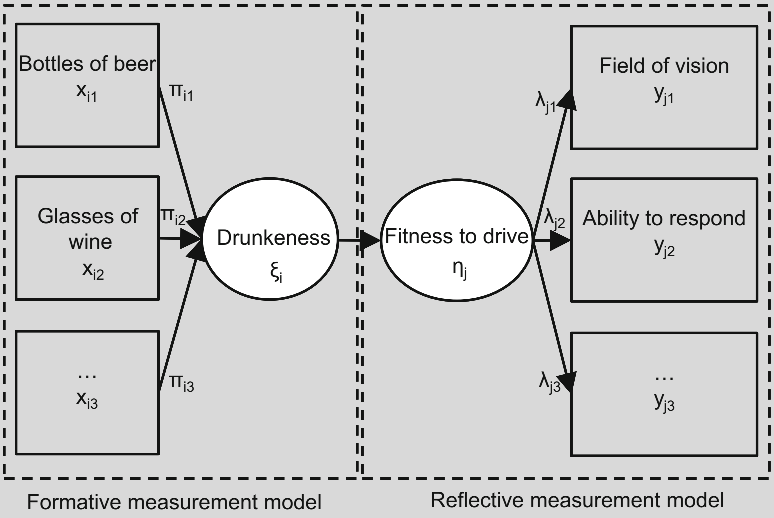


Fig. 7.8 Example of a structural and measurement model (simplified Technology Acceptance Model by Davis et al. 1989—illustration without measurement errors)

"effectiveness", "performance"). This is a perfectly plausible assumption in many social psychological phenomena where it is assumed that an observation (e.g., a verbal opinion) can be explained by an underlying concept: for example, a statement such as, "I like the Apple brand" is "caused" by the attitude to the Apple brand. However, there are also constructs in which the latent variable is explained or caused by the indicators. These indicators are referred to as **formative indicators** (for more detail on reflective vs. formative indicators see Burke et al. 2003).

The difference between formative and reflective indicators can be clearly illustrated by the example of drunkenness and fitness to drive (see Ringle et al. 2006, p. 83). The model is simplified and shows no measurement errors.



The latent variable "drunkenness" is measured by means of formative indicators referring to consumed alcohol, which is the cause of drunkenness. The more that is consumed, the greater the drunkenness. This also shows how important the completeness of the measurement model is. If, for example, only the amount of wine consumed, but not the amount of beer consumed, is measured, the measurement is wrong. Unlike formative ones, for reflective measurement models, the latent variable is the origin of changes in the indicator values. As a result, all the indicators associated with a latent variable are highly correlated, so that the elimination of a single reflective indicator is usually not a problem. In the example, the fitness to drive has an influence both on the size of the field of vision and on the ability to respond.

Structural equation models, in particular the measurement models contained therein, are also often used today to test the **convergent and discriminant validity of measurements** of constructs (see Sect. 6.3.3). On the one hand, the correspondence of several indicators for the measurement of the same construct (→ convergent validity) is tested, and on the other hand, the discriminability of several constructs (→ discriminant validity) are examined (Fornell and Larcker 1981; Hair et al. 2010).

Estimating the parameters of such models requires complex and sophisticated procedures for which appropriate software is available, although, of course, this does not obviate the need for a thorough understanding of the methods to ensure a meaningful application. Software is distinguished into covariance-based techniques (e.g., LISREL / AMOS / MPlus) and variance-based methods (PLS). The result of such an estimation shows whether or not the theoretically suspected relationships between the different variables are confirmed and how strong these relationships are. For such results, so-called fit indices are used to assess the extent to which the theoretical model complies with the data collected. These methodically challenging questions are widely discussed in the literature. For (relatively) easy-to-understand presentations, see Hair et al. (2010).

References

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management, 37*(1), 5–38.

Aguinis, H., Cascio, W. F., & Ramani, R. S. (2017). Science's reproducibility and replicability crisis: International business is not immune. *Journal of International Business Studies, 48*(6), 653–663.

Allison, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge.

Banks, G., O'Boyle, E., Pollack, J., White, C., Batchelor, J., Whelpley, C., et al. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management, 42*(1), 5–20.

Bettis, R. A., Ehtiraj, S., Gambardella, A., Helfat, C., & Mitchell, W. (2016). Creating repeatable cumulative knowledge in strategic management. *Strategic Management Journal, 37*(2), 257–261.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation* (4th ed.). Berlin: Springer.

Burke, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*(2), 199–218.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982–1003.

Eisend, M. (2015). Have we progressed marketing knowledge? A meta-meta-analysis of effect sizes in marketing research. *Journal of Marketing, 79*(3), 23–40.

Ellis, P. D. (2010). *The essential guide to effect sizes: An introduction to statistical power, meta-analysis and the interpretation of research results*. Cambridge: Cambridge University Press.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edingburgh: Oliver & Boyd.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(2), 39–50.

Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). Boston: McGraw-Hill.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hunt, S. D. (2010). *Marketing theory—Foundations, controversy, strategy, resource—Advantage theory*. Armonk, NY: Sharpe.

Jaccard, J., & Becker, M. (2002). *Statistics for the behavioral sciences* (4th ed.). Belmont, CA: Wadsworth.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2), 137–152.

Kerr, N. (1988). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Kruskal, W. (1968). Tests of statistical significance. In D. Sills (Ed.), *International encyclopedia of the social sciences* (pp. 238–250). New York: Macmillan and Free Press.

Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review, 7*(3), 471–479.

Lipsey, M. W., & Wilson, D. T. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Neuman, W. L. (2011). *Social research methods—Qualitative and quantitative approaches* (6th ed.). Boston: Pearson.

Peter, J. (1991). Philosophical tensions in consumer inquiry. In T. Robertson & H. Kassarjian (Eds.), *Handbook of consumer behavior* (pp. 533–547). Englewood Cliffs, NJ: Prentice-Hall.

Ringle, C., Boysen, N., Wende, S., & Will, A. (2006). Messung von Kausalmodellen mit dem Partial-Least-Squares-Verfahren. *Wirtschaftswissenschaftliches Studium, 35*(1), 81–87.

Sawyer, A. G., & Peter, J. P. (1983). The significance of statistical significance tests in marketing research. *Journal of Marketing Research, 20*(2), 122–133.

Selvin, H., & Stuart, A. (1966). Data-dredging procedures in survey analysis. *The American Statistician, 20*(3), 20–23.

Shugan, S. (2002). Marketing science, models, monopoly models, and why we need them. *Marketing Science, 21*(3), 223–228.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1–2.

Further Reading

Ellis, P. D. (2010). *The essential guide to effect sizes: An introduction to statistical power, meta-analysis and the interpretation of research results*. Cambridge: Cambridge University Press.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson.

Jaccard, J., & Becker, M. (2002). *Statistics for the behavioral sciences* (4th ed.). Belmont, CA: Wadsworth.