MARKETING
ZFP – Journal of Research and Management
www.marketing-zfp.de

Research Articles

Avoiding Pitfalls in Experimental Research in Marketing

By Anja Spilski, Andrea Gröppel-Klein and Heribert Gierl

Recently the number of publications and tutorials dealing with how to conduct experiments properly has increased. The present article aims to make novice experimenters aware of important issues that arise when preparing experiments theoretically, conducting them, analyzing the experimental data and interpreting the results. We examine recent literature and provide guidance on how to avoid pitfalls during the experimental process. Regarding theoretical preparation, we discuss different forms of hypotheses (basic form, moderation, mediation, and integrated moderation and mediation), their interpretation, and points to consider when deriving them. Regarding the conducting of experiments, we discuss decisions about the design of variables (manipulation or measurement, realism), samples and ethical issues. Regarding data analysis, we discuss the necessary checks, covariation, dichotomization, effect size and issues relating to the reporting of results. Finally, regarding the interpretation of results, we discuss the non-significance of findings and the generalization of results.

1. Research on experiments in marketing: Relevance and topics

Experimentation is an important methodology in marketing research. A review of the literature by Koschate-Fischer and Schandelmeier (2014) finds that more than 50 % of articles in the four leading marketing journals – *Journal of Marketing*, *Journal of Consumer Research (JCR)*, *Journal of Marketing Research* and *Marketing Science* – are based on experiments. Looking specifically at JCR, Peterson and Umesh (2018) find that 27 % of the first two volumes are experimental studies (1974/75, 1975/76; 15 absolute) while 84 % of the studies published in Volumes 40 and 41 (2013/14, 2014/15) are experiments (151 absolute). Content analyses in other areas of business research (e. g., Bouwman and Grimmelikhuijsen 2016; Fong et al. 2016) also show that experiments are both prevalent and increasing in importance.

There is a significant amount of literature on experimental work. These studies look at the issue of experimentation from different angles. We may categorize them as follows:

(1) Publications asking why and how we carry out experiments in the first place: the fundamental idea behind experimentation.

(2) Guidelines to be followed during the experimental process.

Anja Spilski is Postdoctoral Researcher at the Institute for Consumer & Behavioral Research, Saarland University, Campus A5.4, 66123 Saarbruecken, Germany. Phone: +49 681 302 2135, E-mail: spilski@ikv.uni-saarland.de
*Corresponding author

Andrea Gröppel-Klein is Chair of Marketing and Director of the Institute for Consumer & Behavioral Research, Saarland University, Campus A5.4, 66123 Saarbruecken, Germany. Phone: +49 681 302 2135, E-mail: groeppel-klein@ikv.uni-saarland.de

Heribert Gierl is Professor of Marketing at the University of Augsburg, Universitaetsstrasse 16, 86159 Augsburg, Germany. Phone: +49 821 598 4051, E-mail: heribert.gierl@wiwi.uni-augsburg.de

(3) Statistical methods for analyzing experimental data.

(4) Studies focusing on specific methodological problems.

(5) Warnings against bad practice.

The first group (e. g., Maxwell and Delaney 2004; Shadish et al. 2002) provides a deep understanding of the fundamental idea behind experimentation, experimental design and related terms, such as "causality" and "validity".

The second group (e. g., Geuens and De Pelsmacker 2017; Hsu et al. 2017; Koschate-Fischer and Schandelmeier 2014; Vargas et al. 2017) looks at experiments from a more practical perspective, focusing on questions of how to perform an experiment, what decisions have to be made, and which guidelines should be followed. This approach is especially useful for novices, as experimentation is a very complex issue. To conduct an experiment properly, researchers have to make decisions about a number of basic structural issues, such as the type of hypothesis and the experimental setting. They also need to consider many possible experimental designs, such as between-subjects, within-subjects, and mixed. On top of this, they have to think about details such as manipulation checks, possible confounds, potential consideration of covariates, the size, structure and accessibility of the sample, statistical power and effect size. All of these issues are interrelated, and mistakes in one or more areas may seriously limit the internal or external validity of the results.

The third group (e. g., Field and Hole 2003; Maxwell and Delaney 2004; Tabachnick and Fidell 2014) considers the statistical methods employed for analyzing data and interpreting results, based on statistical tables and figures produced using software such as SPSS.

The fourth group considers specific problems relating to the experimental process. Unlike the second group of studies, these studies are not overviews but more detailed examinations. They include, among others: articles on the use of manipulation checks, e. g., Perdue and Summers 1986; articles on the effects of using convenient samples, e. g., Ashraf and Merunka 2017, Espinosa and Ortinau 2016, Hauser and Schwarz 2016; and articles on the issue of dichotomizing variables, e. g., Iacobucci et al. 2015b.

The fifth group of studies (e. g., Babin et al. 2016; Inman et al. 2018; Pham 2013; Woodside 2016) are critical analyses of the general research process in management, marketing and consumer psychology. As Babin et al. (2016, p. 3133) state, "accepted practices" exist in marketing research that are considered "sacred cows" but which actually "do more harm than good." Pham (2013, p. 411) addresses "seven sins" of research in consumer psychology. Woodside (2016, p. 365) presents a manifesto for overcoming the "bad practices pervasive in current research in business." Several of the issues addressed in these articles also relate to issues in *experimental* research in marketing.

The present article makes two contributions. First, the large number of studies dealing with experimental research might lead to those individuals relatively new to experimental research feeling overwhelmed. The present article provides a starting point for novice experimental researchers. Thus, our article is a synthesis of types 2 (guidelines), 4 (studies of specific problems) and 5 (warnings against bad practice). It makes novice experimental researchers aware of possible pitfalls that may occur during the process, from planning experiments to discussing results. We present a list of questions frequently raised by our own students, from preparing experiments theoretically (3.1) to conducting them (3.2), analyzing the experimental data (3.3) and interpreting the results (3.4). Although it is beyond the scope of the present article to treat all these questions in detail, we provide general answers to certain questions and references to more detailed studies where appropriate.

Second, our review of literature published in the last five years reveals that discussion about experimental procedures and specific techniques is far from over. Thus, Meyvis and Van Osselaer (2018, p. 1157), based on observations by Simmons et al. (2011), state that "[e]xperimental social science, including the field of consumer research, has recently been shaken by what has been termed the 'replication crisis.' Researchers have become aware that many of our field's findings are difficult to replicate and might, in fact, not be true. The estimates of effect sizes in published work tend to be inflated as a result of the use and abuse of 'researcher degrees of freedom,' such as selectively omitting studies or conditions, or making selective decisions about the use of covariates and transformations or about the removal of participants from the analysis." In the present article we draw readers' attention to the recent methodological debates and to the fact that important issues relating to experimentation are still under discussion. We provide an overview of areas subject to current discussion based on a review of literature published in the last five years dealing with issues relating to experiments. Novice researchers are advised to avoid merely citing authors of other experimental studies who use appealing procedures ("they did it that way, therefore so do I") and instead provide arguments in favor of their procedures, drawing attention to recent developments in experimentation.

2. Brief overview of experiments

"One of the beautiful features of experiments is the causal interpretations they afford about differences between groups... [W]hen done well, no research design gives a researcher more confidence in the claim that differences between groups defined by *X* on some variable of interest [*Y*] is due to *X* rather than something else" (Hayes 2018, p. 121). Generally, experiments are used to analyze the relationship between at least two variables: a variable assumed to have an impact (*X*, the independent variable)

and a variable assumed to be impacted (*Y*, the dependent variable).

Typically, independent variables contain levels that are varied by the researchers, while dependent variables are measured. Thus, "[a]n experiment is formed when the researcher manipulates one or more independent variables and measures their effect on one or more dependent variables, while controlling for the effect of extraneous variables" (Malhotra et al. 2017, p. 308). Controlling for extraneous variables makes it possible to say that the independent variables are responsible for the effect, because those independent variables are the only ones that are varied in the experiment. For this reason, experiments can be used to analyze not just correlation-type relationships but also cause-and-effect relationships, where the independent variable impacts on the dependent variable. To investigate the causes of an effect, it is not enough to merely observe or measure certain constructs and then correlate them. It is necessary to systematically vary the conditions of the variable that is assumed to be the cause – a process known as "manipulation" – and systematically measure the consequences of this variation for the dependent variable.

Researchers assign the participants in the experiment either to an experimental group or to a control group. The dependent variable is then measured. In randomized experiments, the researchers assign participants to groups by chance. The idea here is that the participant groups only differ in terms of the conditions represented by the experimental and control groups; it is assumed that the other variables that might vary between participants are more or less evenly distributed between the groups thanks to the random assignment. Researchers can then determine causal relationships between the variables by comparing the different levels of the independent variable with regard to the value of the dependent variable.

Three criteria are considered necessary for causation to be present: covariation of the variables (the cause and the effect must vary together), temporal precedence (the cause must precede the effect), and elimination of competing explanations for the effect (Mill 1843; illustratively explained by Vargas et al. 2017). "Internal validity" may be assumed where manipulation of the independent variable is the only reason for changes in the measured values of the dependent variable. Shadish et al. (2002) provide an extensive list of threats to the internal validity of the observed findings – in other words, reasons for alternative explanations.

To ensure internal validity, the researchers must make many decisions during the experimental process. These decisions usually have consequences for the implications of the results. We illustrate a typical experimental process and the subsidiary questions involved in *Tab. 1*, adding the specific problems that researchers must address. We then use a question-and-answer format to draw readers' attention to certain critical points during the experimental process and reflect on recent discussions in the field.

3. Critical issues during the experimentation process

3.1. Frequently asked questions concerning preparing experiments theoretically

What constitutes a contribution to the marketing literature?

Journal editors and reviewers typically stress the importance of the contribution provided by research (Bagchi et al. 2017; Brown and Dant 2008; Janiszewski et al. 2016; Ortinau 2011). Consequently, the experimental process should start with a research question that is interesting, important and relevant from a scientific and/or practical perspective. Ortinau (2011) provides insight into how such a research question can be found and which environmental factors should be considered when searching for it.

Generally, it is helpful to have an understanding of the type of inquiry the research is. Deductive approaches can usually be distinguished from inductive approaches (Lynch et al. 2012). Owing to our focus on experimentation and hypothesis testing, we consider only deductive approaches in the following. Lynch et al. (2012) distinguish between deductive research categories where the intended contribution is to enhance knowledge in a conceptual versus a substantive domain. The former is interested in theory ("as exemplified by research on dual process models of persuasion, regulatory focus, fluency, construal level," and others; Lynch et al. 2012, p. 475), while the contribution of the latter derives from successful application of existing theoretic constructs to explain observations or analysis of the effectiveness of managerial or public policy interventions (e. g., the role of construal level in advertising effectiveness). It is argued that these types of research should have different bases and criteria when being reviewed (Lynch et al. 2012) and lead to the appropriateness of different research procedures (e. g., research settings and sampling; Calder et al. 1981).

Another way of describing research (from the deductive-conceptual perspective) is the excellent metaphor of knowledge as a "forest of knowledge trees," suggested by Janiszewski et al. (2016). Knowledge creation, for instance, can be the addition of leaves to the tree (more an "incremental innovation" but still a contribution) or the starting of a new branch or the sprouting of a new seedling (a more "radical innovation"). When communicating an article's position to the reviewers, the aspect of contribution can be combined with a reference to the intended "structure" of knowledge creation (Janiszewski et al. 2016).

For the practical issue of how to explain the significance (in a non-statistical sense) of an article's contribution, several authors have provided advice (e. g., Brown and Dant 2008; Ortinau 2011) and stressed that to mention that the research question has not been examined previ-

Step in the experimental process	Typical questions		Selected references (in alphabetical order)
Theoretical preparation of the experiment			
Research question	What constitutes a contribution to the marketing literature?	▪ Types of research ▪ Criteria in the reviewing process	Bagchi et al. 2017; Brown and Dant 2008; Janiszewski et al. 2016; Lynch et al. 2012; Ortinau 2011
Hypotheses	What should I consider when formulating hypotheses?	▪ Form of hypotheses ▪ Types of control groups ▪ Other explanation variables	Armstrong et al. 2001; Woodside 2016
	What are the different approaches to developing hypotheses? Where do hypotheses come from?	▪ Approaches: dominant and competing hypotheses ▪ Consideration of the possibility of multiple theoretical foundations of an effect ▪ Critiques: "studies of theories" and the "mere theories of studies"	Armstrong et al. 2001; McGuire 1997; Pham 2013; Sutton and Staw 1995
	How can I formulate hypotheses about moderation effects?	▪ Difference between independent and moderator variable ▪ Types of interactions ▪ Avoid mixing up arguments on interaction and arguing on the main effect of the moderator ▪ Arguing on the "reverse interaction effect" ▪ Multiple moderators	Andersson et al. 2014; Cohen et al. 2003; Dawson 2014
	How can I formulate hypotheses about mediation effects?	▪ Types of mediation, multiple mediators ▪ Discriminant validity between mediator and dependent variable ▪ Directionality regarding mediator and dependent variable	Aguinis et al. 2017; Demming et al. 2017; Hayes 2018; Pieters 2017; Preacher and Hayes 2008; Voorhees et al. 2016; Zhao et al. 2010
	Can I combine moderation and mediation?	▪ Types of moderated mediation: first-stage, second-stage moderation ▪ Test for differences between conditional indirect effects ▪ Collinearity in moderated mediation	Chennamaneni et al. 2016; Hayes 2018; Pieters 2017
Conducting the experiment			
Experimental design	What should I do if random assignment of participants to conditions is not possible?	▪ Randomized experiment vs. quasi-experiment	Shadish et al. 2002
	Should I use a between-subjects or a within-subjects design?	▪ Extraneous variables: testing effects/carry-over effects, selection bias (among others)	Field and Hole 2003; Charness et al. 2012; Koschate-Fischer and Schandelmeier 2014; Meyvis and Van Osselaer 2018; Shadish et al. 2002
	Should I manipulate or measure variables?	▪ Types of manipulation ▪ Individual difference variables ▪ Series of studies with varying methods ▪ Measurement-of-mediation design vs. moderation-of-process design	Cozby 2005; Geuens and De Pelsmacker 2017; Leary 2012; Perdue and Summers 1986; Spencer et al. 2005
	What should I take into consideration concerning the realism of the experiment?	▪ Types of settings (laboratory, field experiment) and internal and external validity ▪ Realism of the independent variable ▪ Realism of the dependent variable	Geuens and De Pelsmacker 2017; Gneezy 2017; Koschate-Fischer and Schandelmeier 2014; Morales et al. 2017
Measurement	Which items and scales should I use for measurement?	▪ Item selection ▪ Adoption of previously published scales ▪ Adopting vs. adapting measures ▪ Multi-item measurement and single-item measurement ▪ Reversed items ▪ Scale formats	Overview in Geuens and De Pelsmacker 2017. Also Bergkvist and Langner 2017; Bergkvist and Rossiter 2007, 2009; Bruner 2015; Diamantopoulos et al. 2012; Ortinau 2011; Sarstedt et al. 2016
	In which order should I arrange the constructs in the questionnaire?	▪ Dependent and mediator variables ▪ Location of manipulation checks ▪ Location of covariates	Geuens and De Pelsmacker 2017; Kühnen 2010; Meyvis and Van Osselaer 2018; Perdue and Summer 1986
	Must the independent variable be	▪ ANOVA, regression analysis	Hayes 2018; Hayes and Preacher 2014; He et

Tab. 1: Overview of recommended literature to answer questions arising in the process of experimentation

Tab. 1: Overview of recommended literature to answer questions arising in the process of experimentation

Sample	Which sample size is necessary?*	▪ Statistical power ▪ Complexity of designs	Cohen 1992; Faul et al. 2007; Maxwell 2004; Meyvis and Van Osselaer 2018
	What should I take into consideration concerning the use of random samples versus convenience samples?	▪ Types of research and use of random samples ▪ Relevance of sample characteristics ▪ Student samples ▪ Crowdsourcing samples (e. g., MTurk)	Ashraf and Merunka 2017; Calder et al. 1981; Espinoza and Ortinau 2016; Goodman and Paolacci 2017; Peterson 2001; Peterson and Merunka 2014; Sarstedt et al. 2017; Shank 2016; Wessling et al. 2017
Ethics	Should I tell respondents what the study is about?	▪ Ethics, informed consent ▪ Demand artifacts ▪ Debriefing	Allen 2004; Geuens and De Pelsmacker 2017; Leary 2012; Meyvis and Van Osselaer 2018; Sawyer 1975; Vargas et al. 2017
Data analysis			
Data analysis methods	Which statistical methods are required to analyze experimental data?*	▪ Analysis of variance in different forms ▪ Analysis of covariance ▪ Regression analysis	Field 2018; Hayes 2018
Checks	Which checks should I consider before testing the hypothesis? Can I remove cases if those checks indicate lower quality?	▪ Manipulation checks ▪ Confounding checks ▪ Quality checks ▪ Instructional manipulation checks	Berinsky et al. 2014; Geuens and De Pelsmacker 2017; Meyvis and Van Osselaer 2018; Oppenheimer et al. 2009; Perdue and Summers 1986; Shadish et al. 2002
Covariation	When should I consider covariates?	▪ Correlation of covariate and dependent variable ▪ Measurement of covariates ▪ Goals of statistical power or adjustment ▪ Correlation of (vs. causal relationships between) covariate and independent variables	Field 2018; Meyvis and Van Osselaer 2018; Miller and Chapman 2001; Tabachnick and Fidell 2014; Yzerbyt et al. 2004
Dichotomization	I have considered a moderator variable that has been measured as a continuous variable. Can I dichotomize in order to calculate ANOVA?	▪ Threats: potential spurious findings, loss of power ▪ Relevance of multicollinearity between constructs ▪ Ongoing discussion	Fitzsimons 2008; Iacobucci et al. 2015a, 2015b; Irwin and McClelland 2003; MacCallum et al. 2002; McClelland et al. 2015; Rucker et al. 2015; Spiller et al. 2013
Significance and effect size	Why should I include an effect size calculation?	▪ Statistically significant findings vs. scientifically significant findings ▪ Effect size vs. importance of effect	Eisend 2015; Hayes 2018; Kline 2013; Lachowicz et al. (in press); Preacher and Kelley 2011; Prentice and Miller 1992
Reporting	Which parameters of the findings should be reported?	▪ Structure of reporting ▪ Typical reporting errors ▪ Necessary parameters for replications ▪ Call for disclosure of non-significant results	APA 2010; Babin et al. 2016; Bakker and Wicherts 2011; Field and Hole 2003; Lehmann and Bengart 2016; Ortinau 2011; Tabachnick and Fidell 2014
Interpretation of the findings			
Rejection of hypotheses	What should I do if the effect that I proposed turns out to be not what I expected?	▪ Non-significant effects ▪ Significantly non-significant effects ▪ Surprising findings ▪ Research ethics	Armstrong 2003; Babin et al. 2016; Banks et al. 2016; Peterson and Umesh 2018
Generalization	What should I take into consideration concerning the generalizability of the results?	▪ Generalizing from the sample drawn to the population ▪ Generalizing from the context used to other possible contexts ▪ Generalizing from lab to field or from non-behavioral measures to behavioral measures ▪ Call for replications: direct, conceptual, with extensions	Calder et al. 1982; Crandall and Sherman, 2016; Lynch et al. 2015; Morales et al. 2017; Uncles and Kwok 2013

Notes: * refers to questions that are not explained in the following sections but included here to propose references for further reading.

Ways to make significant contributions	Explanations
1. Adding new knowledge	a) Applying new theories to existing problems b) Filling in knowledge gaps c) Investigating antecedent variables heretofore overlooked d) Studying consequent variables heretofore ignored e) Examining overlooked intervening or mediating variables
2. Deepening our understanding of existing knowledge	a) Identifying a theory's boundary conditions: i) examining potential moderator effects ii) probing the theory's external validity iii) testing the theory's assumptions b) Reconciling contradictory findings
3. Uncovering surprising results	"Results that challenge conventional wisdom about theoretical linkages that are believed to be written in stone"
4. Tackling problems that interest practitioners	Research and findings that challenge "conventional managerial practices or beliefs"

Tab. 2: Ways to make significant contributions according to Brown and Dant (2008, p. 134)

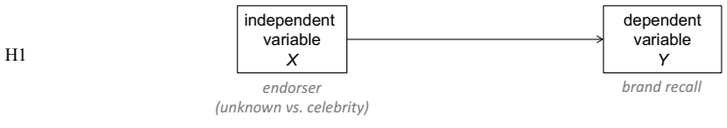

Fig. 1: Illustration of a basic hypothesis

ously is a weak positioning argument and not enough for a significant contribution. Research questions that have not been considered in the past may be unimportant research questions (Varadarajan 1996). In addition, claims that a study is the first to analyze a particular research question in a specific country may not provide a sufficient contribution (Geuens and De Pelsmacker 2017).

Brown and Dant's (2008) definition in achieving contribution may also be helpful here. They focused on contributions to retail research and defined a significant contribution to the retailing literature "as research that tackles interesting and relevant retailing-related issues, advances our theoretical and/or methodological understanding of those issues, and deepens our knowledge of those issues" (Brown and Dant 2008, p. 132). This understanding of a contribution can be transferred to marketing-related or consumer-related research areas other than retailing research (see similarly Morales et al. 2017). Brown and Dant (2008) differentiate four ways to make significant contributions (*Tab. 2*). Experimental analysis is able, for instance, to consider moderation of direct or mediated effects, but can also yield surprising results and thus also challenge practitioners' beliefs.

What should I consider when formulating hypotheses?

Experimental research requires thinking in cause and effect. Typically, these cause-and-effect relationships are summarized by hypotheses. The hypothesis, in its verbal form, should summarize the research question as a relationship between the independent and dependent variables. A directed hypothesis also states whether the effect is positive or negative (i. e., which condition of the independent variable will result in the higher value of the dependent variable[1]).

Imagine the following example taken from research into celebrity endorsement effects (Erfgen et al. 2015): Researchers are interested in whether the use of a celebrity endorser to advertise a branded product would lead to an effect of overshadowing the advertised brand ("vampire effect"), with the result that "consumers remember only the celebrity, not the brand" (Erfgen et al. 2015, p. 155). A hypothesis can be set in the format of a proposition, as demonstrated by Erfgen et al. (2015, p. 156):

H1: Recall of the brand is lower when the advertisement contains a celebrity endorser than when it contains an equally attractive but unknown endorser.

Alternatively, hypotheses can come in the format of if-then statements (Sekaran and Bougie 2016): *If the advertisement contains a celebrity endorser, then recall of the brand will be lower than if it contains an equally attractive but unknown endorser.*

Although both formats contain identical meaning, the if-then format can be considered as more intuitive, since it represents the causal order, starting the statement with the independent variable. This causal order often is also illustrated by path diagrams (Hayes 2018), which we will use for the basic hypothesis ($X \rightarrow Y$, see *Fig. 1*) and the other types of hypotheses described in the following sections.

Experimental research requires critical thinking to rule out plausible alternative explanations for the effect. In the celebrity endorser example, Erfgen et al. (2015) iden-

tified attractiveness of the endorser as a potential further explanation of a vampire effect: "Because attractiveness has a strong positive impact on brand recall ... it is not clear whether the higher recall values for celebrities were due to their celebrity status or their higher attractiveness" (p. 156). In their hypothesis, Erfgen et al. (2015) rule out this potential explanation by ensuring constancy of attractiveness across the conditions.

Of particular relevance is that the control group is also specified in the hypothesis. In this context, active and passive control groups can be distinguished. Woodside (2016), more illustratively, referred to the different types as "placebo" and "nocebo" control groups. An active (placebo) control group would also receive a stimulus, which "looks to be the same experience but without the actual substance expected to cause the focal outcome as part of the experience" (Woodside 2016, p. 377). A passive (nocebo) control group simply would not receive a stimulus. Woodside (2016) generally recommends administering placebo groups. In an experiment, the comparison of a treatment group (presenting a stimulus) with a nocebo group (no stimulus) might result in bias because of reactivity effects (individuals alter their behavior simply because they are aware that they are being studied). In the Erfgen et al. (2015) example, a nocebo control group would not make sense since, for measuring recall, both groups have to be presented with a stimulus. However, there are situations in which nocebo groups could be included, for example, in a study measuring a treatment group's attitude towards a brand receiving negative publicity (treatment) and comparing it to a group of participants where attitudes are measured without receiving any information about the brand. In cases like this, Woodside (2016, p. 377) recommends "using three groups in a true experiment to include treatment, placebo, and nocebo conditions in order to examine the effects of experiencing the giving/receiving the administration steps that occur in the study versus not experiencing these steps. This three group design could test the hypothesis that the responses by the participants in the nocebo group were lower than responses of participants in the placebo group which were lower than the responses by participants in the treatment group."

What are the different approaches to developing hypotheses? Where do hypotheses come from?

Armstrong et al. (2001) distinguish dominant and competing hypotheses. Dominant hypotheses are based on a certain theory. Armstrong et al. (2001, p. 173) suggest that a "dominant hypothesis, designed to rule out a null hypothesis, often becomes a search for evidence to support a favored hypothesis" – referred to as "advocacy hypothesis construction and testing" by Woodside (2016, p. 378). The authors warn that researchers can become subject to a "confirmation bias" that makes them blind to other explanations for a specific effect.

Competing hypotheses are used to analyze which of a variety of theories can be considered to best describe the underlying process of the effect (rival theories). Therefore, two or more plausible hypotheses are tested (Armstrong et al. 2001). The use of competing hypotheses "enhances objectivity because the role of the scientist is changed from advocating a single hypothesis to evaluating which of a number of competing hypotheses is best" (Armstrong et al. 2001, p. 175). In *Tab. 3*, an example (the issue of presenting advertisements in a media context) is provided that contains dominant and competing hypotheses.

In a publication audit conducted in six marketing journals for the time period 1984 to 1999, Armstrong et al. (2001) found that the dominant hypothesis approach featured in the majority of studies: 74.4 % of 1,701 studies contained a dominant hypothesis; only 13 % contained competing hypotheses. These figures relate to articles using diverse methodologies; the share of articles applying experimental design and competing hypotheses would probably be even lower. In more recent critiques of research practices in marketing, however, scientists call for increased use of competing hypotheses that allow the testing of rival theories (Woodside 2016). Future use of competing hypotheses can also be considered from the perspective of the availability of previous findings: As knowledge in marketing science increases, the number of theoretical approaches grows and marketing scientists increasingly use interdisciplinary approaches (Kenworthy and Sparks 2016), we could assume that more research problems will demand the consideration of competing

Types of hypotheses	Examples from the area of advertising in a media context
Dominant hypothesis: Hypothesis assuming a cause-and-effect relationship with a particular prediction on the direction	Inducing positive (vs. no) emotions by a media context will result in a more positive evaluation of an advertisement that is embedded in that context.
Competing hypotheses: Hypotheses assuming a cause-and-effect relationship with different predictions on the underlying theory of the effect ("rival theories")	**Theory 1:** Commercials (irrespective of which emotional tone they have) will be more effective when they are embedded in *happy* program content: Hypothesis 1: "[C]ognitive responses, commercial evaluation, and purchase intention will be significantly more favorable for consumers viewing upbeat or sad commercials presented in the context of happy program content than for consumers exposed to identical commercials in the context of a sad program." (Kamins et al. 1991, p. 3) **Theory 2:** Commercials will be more effective when they are embedded in *matching* program context, e. g., sad commercials will be effective when they are embedded in sad program context. Hypothesis 2: "[C]ognitive responses, commercial evaluation, and purchase intention will be significantly more favorable for consumers exposed to a commercial which matches the affective tone of the program content in which it is embedded." (Kamins et al. 1991, p. 5)

Tab. 3: Examples of types of hypotheses

hypotheses. However, as Armstrong et al. (2001, p. 181) assume, papers containing competing hypotheses "are seldom published" because they are "more likely to produce controversial findings" and "challenge existing theories." Therefore, calls for more frequent use and acceptance of competing hypotheses are not only directed to authors, but also to journal editors and reviewers (Armstrong et al. 2001; Woodside 2016).

However, Pham (2013) stresses that it is unlikely that effects as complex as those analyzed in consumer behavior research and consumer psychology can be based on "the *single* best explanation" (p. 415), and calls for greater consideration of the possibility of multiple theoretical foundations of an effect. From this perspective, a decision for one of two or more competing hypotheses could simply not say enough, as "many theories should be seen as complementary rather than competing because they capture different levels of explanation" (Pham 2013, p. 415). McGuire (1997) recommends "to test not *if* a given explanation does or does not account for a significant proportion of the variance in the hypothesized relation, but to test *to what extent* the relation is accounted for by each of several explanations" (p. 17, emphasis added).

When should researchers use which approach? Armstrong et al. (2001) recommend considering the amount of available prior knowledge concerning the research questions. An exploratory approach is used when there is little prior research, the dominant hypothesis approach when prior knowledge is to be refined and extended by analyzing boundary conditions, and the competing hypotheses approach where prior knowledge leads to two or more reasonable explanations.

Theory is the basis of developing arguments for hypotheses. Researchers should present what the particular theory is about and discuss the logical arguments as to why this theory has led to the predictions. Sutton and Staw (1995) provide guidance in "what theory is not" in order to assist authors regarding adequate argumentation. They claim that "references, data, variables, diagrams, and hypotheses are not theory" (Sutton and Staw 1995, p. 371). They emphasize that (1) it is not enough to merely list prior research but necessary to set prior research in the context of the own research question; (2) it is not enough to argue that others have reported data on certain findings and that similar patterns would be expected from the data, but it is necessary to transfer prior research findings to the new argumentation; and (3) it is not enough to list and define variables but necessary to explain the relationships between these variables. Moreover, researchers should not confuse comparative tests of variables with comparative tests of theory. Further, Sutton and Staw (1995) stress that (4) presenting diagrams that show relationships in a logical ordering (although this is helpful) and (5) a mere listing of hypotheses cannot substitute for a set of logical explanations.

A further critique is raised by Pham (2013, p. 420), who makes a distinction between "studies of theories" and "mere theories of studies." The former addresses the desirable testing of novel theoretical propositions, whereas the latter refers to studies that do not model real-world phenomena but include "the conceptualization of a very narrow phenomenon that most likely only occurs under the artificial conditions that the researchers seek to create in the lab" (p. 421).

How can I formulate hypotheses about moderation effects?

Hypotheses on moderation effects determine under which conditions an effect will occur. In other words, we can "better understand some phenomenon when we can answer not only whether *X* affects *Y*, but also ... *when X* affects *Y* and when it does not" (Hayes 2018, p. 6).

A moderator could also predict a strengthening or weakening of the $X \rightarrow Y$ relationship under different moderator conditions. In sum, the moderating variable influences the independent variable-dependent variable relationship. Moderation is typically illustrated as shown in *Fig. 2* panel B, which shows two arrows – one representing the $X \rightarrow Y$ path and the other representing the influence of the moderator on the $X \rightarrow Y$ path.

Reconsider the celebrity endorsement example by Erfgen et al. (2015, p. 157) where a moderator hypothesis is formulated based on the basic hypothesis already shown in *H1*:

H_{Mod}: *The negative effect of a celebrity endorser on brand recall is greater (lesser) in conditions of low (high) brand familiarity.*

In this hypothesis, the moderator variable contains conditions of low vs. high brand familiarity. The moderation hypothesis predicts that the effect of the absence vs. presence of a celebrity endorser (*X*) on brand recall (*Y*) is contingent on whether or not consumers already know the brand. When consumers *do* already know the brand, the effect is proposed to be negative; when they do *not* already know the brand, the effect is proposed to be *even more* negative. In statistical terms, two contingent effects can be tested: an effect of *X* on *Y* under conditions of low brand familiarity and an effect of *X* on *Y* under conditions of high brand familiarity. The effect of the moderator on the $X \rightarrow Y$ relationship is referred to as an interaction effect that can be tested for statistical significance.

The difference between an independent variable and a moderator variable is in which part of the model they exert influence. The independent variable affects the dependent variable ($X \rightarrow Y$); in the example, absence vs. presence of a celebrity endorser is supposed to affect brand recall. If a further variable is supposed to be a predictor of *Y*, then this variable is also an independent variable. This is imaginable for prior brand familiarity, which may have an influence on brand recall. If we combine both effects in one model, we would have two independent variables, X_1 = presence vs. absence of a celebrity endorser, X_2 = brand familiarity, which could both in-

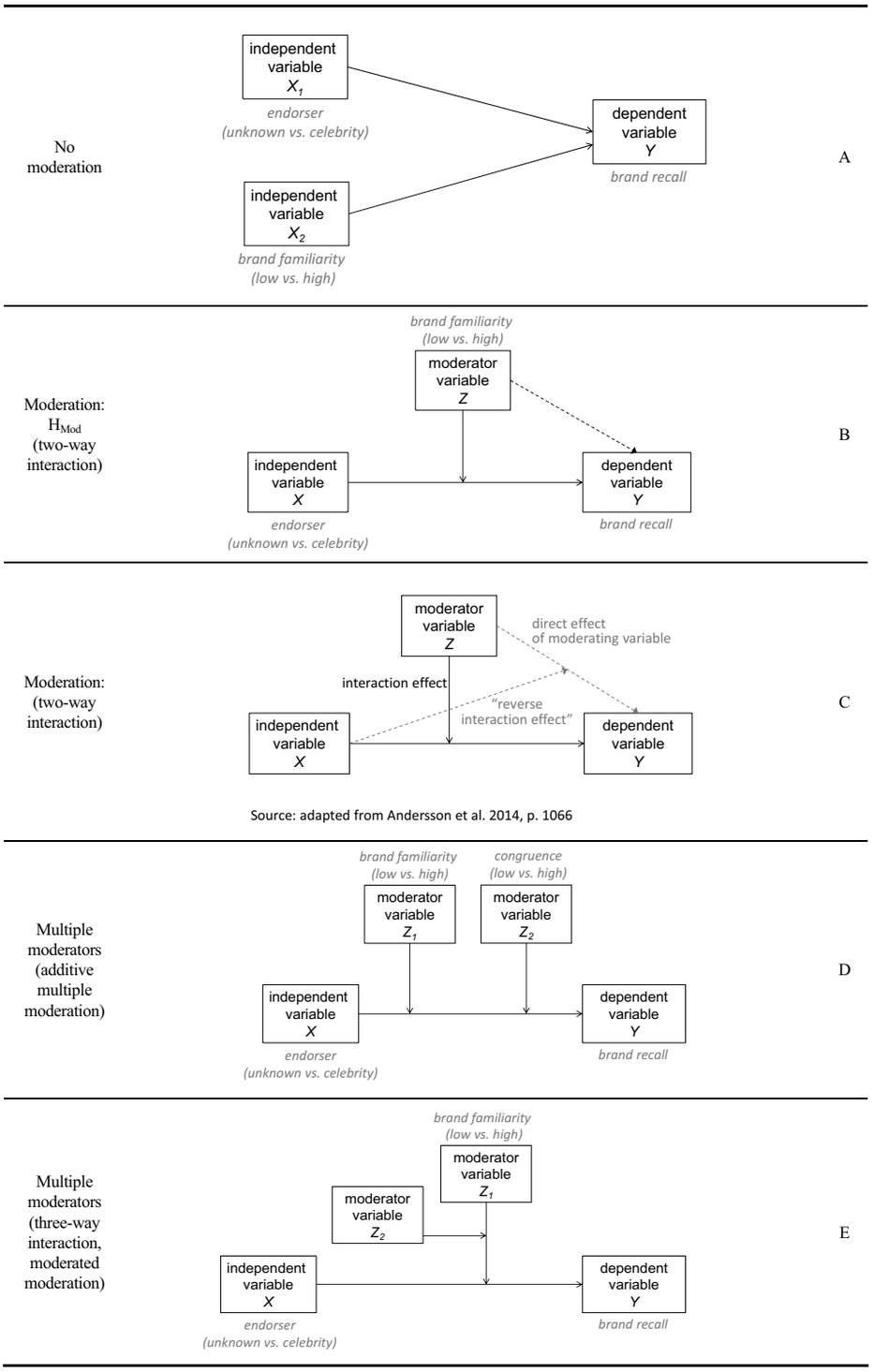

Fig. 2: Illustration of non-moderation and different forms of moderation

dependently influence the dependent variable, Y = brand recall. In this case, X_2 would not be a moderator (see *Fig. 2*, panel A). Alternatively, X_2 can be considered a covariate. In contrast, when it is supposed that X_2 has an influence *on the effect* of X_1 on Y, this variable becomes a moderator (see *Fig. 2*, panel B). In the following, we refer to moderator variables as Z.

Moderation hypotheses represent contingency hypotheses. They are used to examine the boundaries of a theory and specify the conditions under which a given theory applies or does not apply (Andersson et al. 2014). In order to derive moderation effects theoretically, Andersson et al. (2014) offer some recommendations for the "ingredients" of the theoretical argumentation, which are briefly summarized here:

- The "independent variable and the moderator variables should not be theoretically related as this would imply mediation" (Andersson et al. 2014, p. 1065).

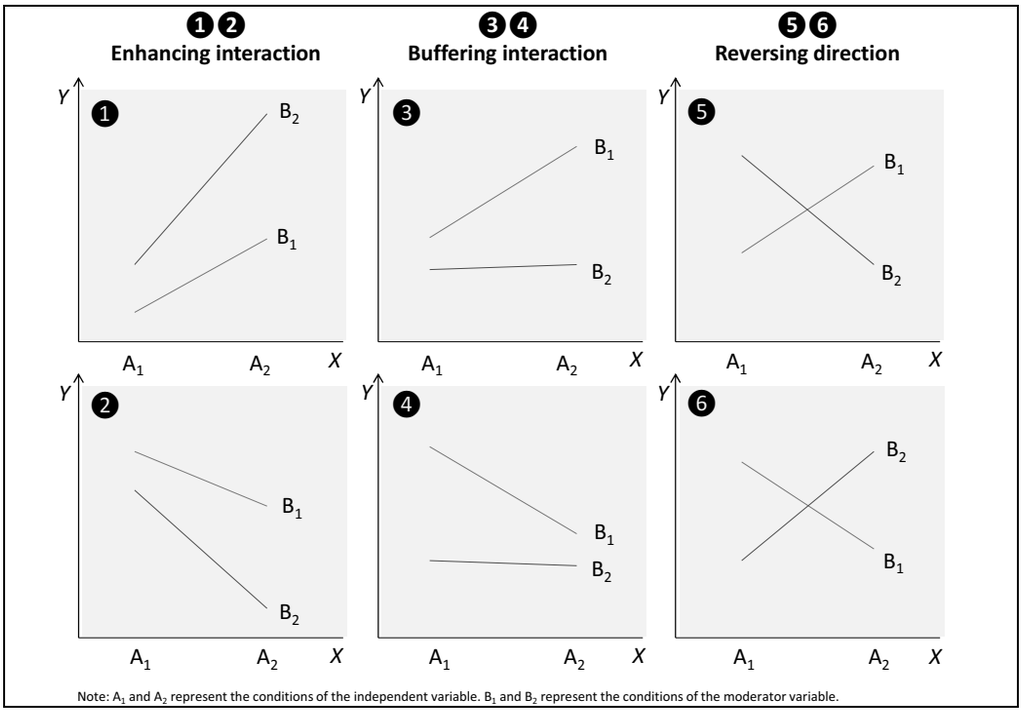

Fig. 3: Patterns of two-way interaction

- If possible, the researchers should first explain *theoretically* the *X*⟶*Y* relationship. Without this knowledge, it would remain "unclear what baseline effect the interaction is supposed to modify" (Andersson et al. 2014, p. 1065). This is particularly relevant when several theories can be used to explain the *X*⟶*Y* relationship because a particular moderator *Z* could only make sense from the perspective of a theory A but not theory B. However, in cases where an interaction effect exists, the first-order effect (main effects of *X*) should not be interpreted, because the interaction effect indicates that *X* does "not have a simple uniform relationship" (Aguinis et al. 2017, p. 672) with Y but that the effect depends on the levels of the moderator variable.

- The model used to test interactions typically contains three terms: the effect of *X* on *Y*, the effect of the moderator *Z* on *Y*, and the interaction effect of *X* and the moderator *Z* on *Y* (represented by a product term *X*×Z). In statistical terms, this interaction effect is symmetrical (Aiken et al. 1991, p. 10). It can be interpreted from two perspectives: First, the effect of *X* on *Y* is contingent on the moderator variable. Second, the effect of the moderator variable on the dependent variable is contingent on *X* (Dawson 2014). Andersson et al. (2014, p. 1066) refer to this second interpretation as the "reverse interaction effect" (see *Fig. 2*, panel C). The interaction will show the same values regardless of which perspective is taken. Andersson et al. (2014) recommend ruling out the reverse interaction in which the independent variable becomes the moderator. This would only be relevant if a theoretical rational exists for directly linking the moderator to *Y*. "The theoretical challenge is to argue that the moderation can only exist in one direction and not the other, for example, because the moderator operates at a different level of analysis or temporally precedes the relationship" (Andersson et al. 2014, p. 1067). Avoiding mixing the perspectives of the interactions is also relevant for graphing the interactions: the effect is illustrated separately for the different conditions of the moderator.

- A frequent mistake is that instead of arguing on the interaction effect, arguments are given for the main effect of the moderator variable on the dependent variable (see *Fig. 2*, panel C). The arguments for a variable's moderating effect on the *X*⟶*Y* relationship must be distinct from the arguments for the moderator's direct effect on *Y* (Andersson et al. 2014).

A moderation effect can come in different patterns; more specifically, the moderator can alter the strength or nature of the effect of the independent variable on the dependent variable. The researchers are recommended to not only hypothesize the existence of an interaction effect, but also its form (Dawson 2014):

- The consideration of the moderator can lead to an increase (or decrease) of a positive or negative *X*⟶*Y* relationship (see *Fig. 3*, panels 1 and 2). Both the independent variable and the moderator variable affect the dependent variable "in the same direction, and together they produce a stronger than additive effect on the outcome" (Cohen et al. 2003, p. 285). Note that in the later test, it is necessary that the conditional effects are statistically significant from each other, indicated by a statistically significant interaction effect (Dawson 2014).

- A further pattern is represented by the case where the moderating variable weakens the effect of the independent variable (Cohen et al. 2003). An example is when there is an $X \rightarrow Y$ effect in one condition of the moderator variable and this effect is absent in the second condition of the moderator variable (see *Fig. 3*, panels 3 and 4). Note that it is not enough that one of the conditional effects is statistically significant while the other is not (Pieters 2017, p. 708). In the later test, the interaction effect has to be statistically significant in order to speak of moderation.
- Moderation can also describe a reversal in the effect's direction determined by a moderator variable. In this case, there would be a positive effect in one condition of the moderator and a negative effect in the second condition (see *Fig. 3*, panels 5 and 6). Again, the interaction effect has to be statistically significant in order to declare it moderation.

Multiple moderators: More than one moderator can be considered in the same model. Besides brand familiarity as a moderator (see example above), Erfgen et al. (2015) considered further moderators of the effect of the use of a celebrity endorser on brand recall. For example, they also considered the congruence of the endorser and the endorsed brand, represented by the conditions of low congruence and high congruence. Here, the moderation suggests that the effect of the absence vs. presence of a celebrity endorser (*X*) on brand recall (*Y*) is contingent on whether consumers perceive low or high congruency. The negative $X \rightarrow Y$ effect becomes stronger when there is low perceived congruence compared to high perceived congruence (Erfgen et al. 2015). The two examples of moderation (by brand familiarity, by congruence) represent separate two-way interactions that can be combined in one model (referred to as "additive multiple moderation", Hayes 2018, p. 320).[2] Brand familiarity is proposed to be one moderator on the celebrity endorser-brand recall relationship, while congruence is another. This can be illustrated by two moderator paths directed on the $X \rightarrow Y$ relationship (see *Fig. 2*, panel D).

There is also the possibility of integrating a further variable (a second moderator) that influences the two-way interaction. This is referred to as three-way interaction (or "moderated moderation", Hayes 2018, p. 320). It describes the presumption that any pattern of the initial interaction effect (as illustrated in *Fig. 3*: enhancing, buffering, reversing direction of an effect) varies across the levels of a second moderator variable (see *Fig. 2*, panel E). Further moderators can be included as well (resulting in even higher-order interaction).

From a methodological perspective, the inclusion of moderating variables increases the number of variables that have to be varied systematically, and it increases the number of "cells" and therefore the sample size necessary to test for effects. Researchers warn against designing experiments that contain "numerous (often minor or subtle) cues" that "might not be attended to, comprehended by, or cognitively processed by the research subjects" (Peterson and Umesh 2018, p. 85). They see this threat especially relevant in three-way, four-way, or higher-level interaction manipulations. Too many moderators, and therefore too many boundary conditions, for an effect might also raise questions of the practical relevance of the findings (Babin et al. 2016, p. 3137). This is not to say that researchers ought to avoid higher-order interaction hypotheses; however, they should critically assess the practical relevance of the results and present them in a comprehensible way. It is not enough to state that an interaction has been found; the interaction has to be interpreted, and its substantive meaning needs to be explained from a theoretical perspective (Andersson et al. 2014).

How can I formulate mediation hypotheses?

So far we have considered the questions of *whether* (at all) and *when* (moderation) the independent variable influences an outcome. Additionally, experimental research is interested in the examination of the *underlying processes* that causally link *X* to *Y* (Hayes 2018). This is considered in mediation analysis. Mediators (*M*) "are conceptualized as the mechanism through which *X* influences *Y*. That is, variation in *X* causes variation in one or more mediators *M*, which in turn causes variation in *Y*" (Hayes 2018, p. 7).

Imagine another example from celebrity endorsement research. Researchers would like to find out whether congruence (vs. incongruence) between the celebrity endorser and the product influences purchase intention of the endorsed brand (similarly Kamins and Gupta 1994; Till and Busler 1998). The basic hypothesis could be as follows: *If there is high congruence between the celebrity endorser and the endorsed brand, then intention to purchase the brand will be more positive than in case of low congruence*. This hypothesis specifies the independent variable *X* and the dependent variable *Y*. However, if the researchers were also interested in the underlying process of this relationship, they would consider a mediating variable. A potential mediating variable in this example could be perceived trustworthiness of the celebrity endorser. The researchers then would assume that variation in celebrity endorser-product congruence (low vs. high, *X)* causes variation in perceived trustworthiness of the endorser (*M)*, which in turn causes variation in intention to purchase the endorsed brand (*Y).* The hypothesis is typically formulated as follows:

H_{Med}: *If there is high congruence between the celebrity endorser and the endorsed brand, then intention to purchase the brand will be more positive than in the case of low congruence. This effect is mediated by perceived trustworthiness of the endorser.*

In statistical terms, in the basic mediation model (see *Fig. 4*, panel A), the total treatment effect on *Y* is decomposed into an indirect effect and a direct effect (the remaining treatment effect, referred to as conditional direct

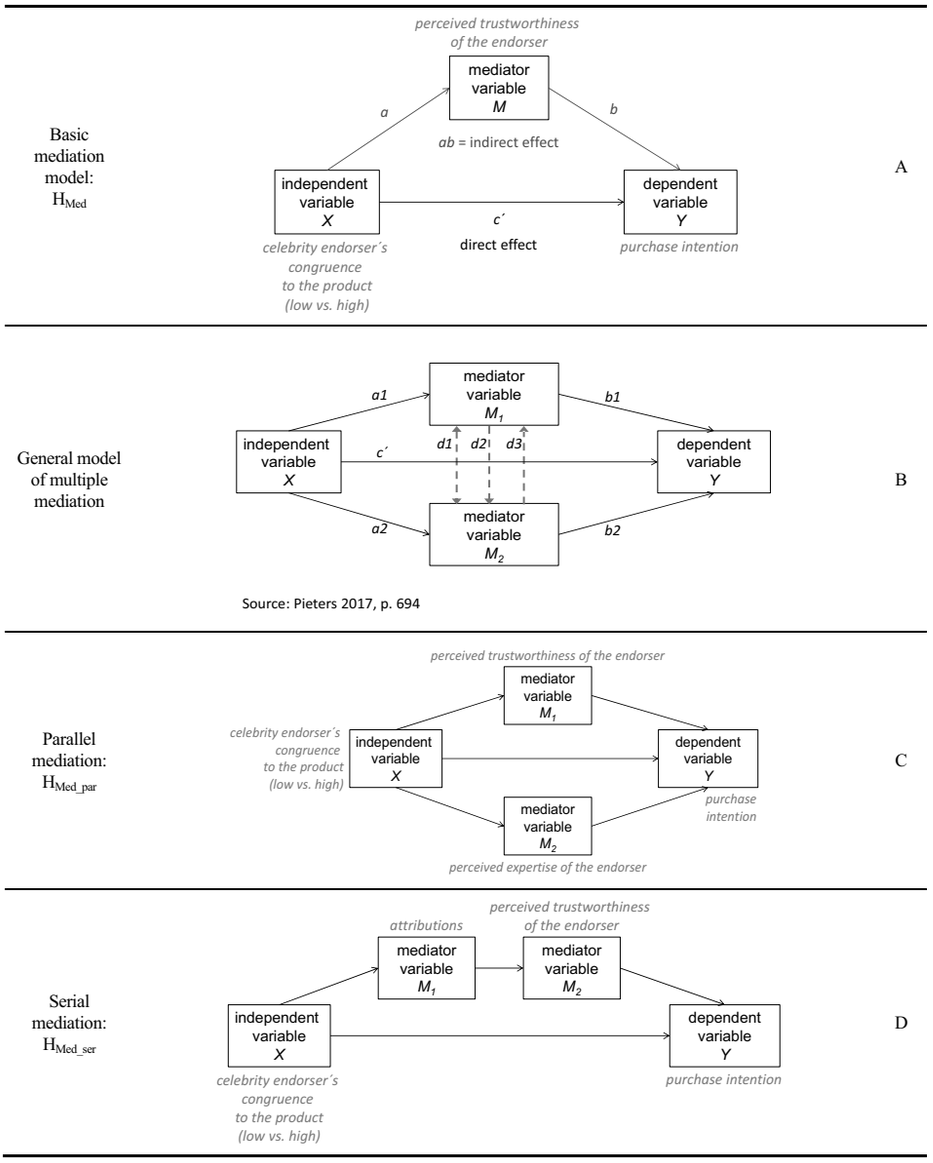

Fig. 4: Illustration of different forms of mediation

effect, *c'*, by Pieters 2017). These are two pathways by which the independent variable is proposed to influence the dependent variable. The pathway leading from *X* to *Y* without passing through the mediator *M* is called the direct effect of *X* on *Y*. The second pathway describes the indirect effect of *X* on *Y* through *M*. The analytical goal is to estimate the regression coefficients *a*, *b*, and *c'* and to interpret them. Coefficient *a* estimates the effect of *X* on *M*. Coefficient *b* quantifies the influence of *M* on *Y* while controlling for *X* (Hayes 2018). Multiplying the coefficients *a* and *b* (resulting in the product *ab*) yields the size of the indirect effect: here, the indirect effect of congruency on purchase intention through endorser trustworthiness. "The indirect effect tells us that two cases that differ by one unit on *X* are estimated to differ by *ab* units on *Y* as a result of the effect of *X* on *M* which, in turn, affects *Y*" (Hayes 2018, p. 84). It is important to consider the signs of *a*, *b*, and *ab*. A hypothesis that predicts a positive indirect effect based on proposed positive *a* and *b* coefficients cannot be supported if the estimated indirect effect is positive but formed by two negative coefficients (Hayes 2018). Preacher and Hayes (2008), Hayes (2018), Pieters (2017), Aguinis et al. (2017), and Demming et al. (2017) provide guidance in mediation analysis.

For models with more than one measured variable (mediator and dependent variable), a concern recently mentioned in the literature refers to the discriminant validity between the mediator variable and the dependent variable. Voorhees et al. (2016, p. 120) state, "for experimental studies that make predictions about the effect of a manipulated independent variable on mediator and dependent latent constructs, a lack of discriminant validity calls into question whether a significant mediator-dependent variable path is just an empirical artifact or measuring the same variable twice." Basically, as early as the conceptual stage of experimentation, the researchers should be aware of the issue of discriminant validity be-

tween mediator and dependent variable and "examine conceptually distant Ms and Ys" (Pieters 2017, p. 697). In the example above, one would assume trustworthiness of the endorser and purchase intention towards the brand as conceptually distinct constructs. However, there are calls for stronger consideration of *testing* for discriminant validity in experiments (Pieters 2017; Voorhees et al. 2016). In a review of articles published in seven leading marketing journals from 1996 to 2012, Voorhees et al. (2016) found that the majority of studies lacked testing for discriminant validity. For experimental studies, this lack was especially pronounced, leading Voorhees et al. (2016, p. 131) to the assumption that "this is likely due to a misconception that discriminant validity testing is something that can only be tested in an SEM context." They discuss and recommend appropriate techniques to test for discriminant validity in experiments.[3] For cases where these tests do not provide evidence in favor of discriminant validity, Pieters (2017) recommends refraining from statistical mediation analysis. Instead, "the mediator and outcome measures are more properly treated as substitute measures of a single outcome" (Pieters 2017, p. 701).

A second concern for models with more than one measured variable is causal directionality (Pieters 2017).[4] The common procedure is that mediator and outcome measures are collected in a single experimental session (Pieters 2017). Even if the mediator and the dependent variable prove to have discriminant validity, the researchers have to provide theoretical reasoning that the most plausible causal direction of influence is from the mediator to the dependent variable ($M \rightarrow Y$) (Pieters 2017) and to rule out that the dependent variable influences the mediator ($Y \rightarrow M$) or that they have a merely correlational relationship ($M \leftrightarrow Y$). As Pieters (2017) states, it is not possible to base this argumentation on statistical grounds but on "logic, theory, and prior research findings" (Pieters 2017, p. 698). In the example above, one would assume that the logic of a causal chain is represented by an effect of the endorser-product congruence on the trustworthiness of the endorser first, which in turn influences intention to purchase the brand. As an alternative to conceptual reasoning, causal directions are proposed to be more effectively analyzed by a series of experimental designs than by mediation analysis. Such a series of studies would (1) manipulate the independent variable and analyze its effect on the measured mediator and dependent variable; (2) manipulate the mediator and analyze its effect on the measured dependent variable (Pieters 2017, see Imai et al. 2013); and (3) manipulate the mediator and analyze its effect on the relationship between the independent and the dependent variable (Spencer et al. 2005). However, this approach relies on the availability of manipulation procedures for the mediator (see section 3.2 "Should I manipulate or measure variables?").

Multiple mediators: The analysis of mediating effects allows the consideration of more than one mediator, operating either in parallel or sequentially. Voorhees et al. (2016, p. 131) observe that marketing "models are going to be exposed to greater refinement and extension." It can be assumed that this is also achieved by models that increasingly consider mediators acting in parallel or serially.

The issue of *parallel* mediators concerns the question of whether several processes exist that mediate the relationship between independent and dependent variables. Reconsidering the example above, variation in congruence of the celebrity endorser and the product may influence the intention to purchase the brand not only via perceived endorser trustworthiness, but also via perceived endorser expertise. Both perceived trustworthiness and expertise of the endorser would serve as mediators, assumed here to act in parallel (see *Fig. 4*, panel C). In this case, the researchers would derive the following hypothesis:

H_{Med_par}: *If there is high congruence between the celebrity endorser and the endorsed brand, then intention to purchase the brand will be more positive than in the case of low congruence. This effect is proposed to be mediated in parallel through perceived trustworthiness and perceived expertise of the endorser.*

Owing to more than one mediator in the model, several "specific indirect effects" (Hayes 2018, p. 152) are hypothesized and estimated later. In this case, one specific indirect effect refers to the influence of congruence on purchase intention through perceived trustworthiness (*M1*) and the other to the influence of congruence on purchase intention through perceived expertise (*M2*).

An assumption of parallel mediation is that, although mediators are potentially correlated (Hayes 2018, Pieters 2017)[5], no mediator causally influences another mediator. Pieters (2017, p. 693) addresses the three possibilities of correlation between different mediators which he referred to as the "*d*-link." In parallel mediator models, the *d*-link of *M1* and *M2* is undirected (see *d1* in *Fig. 4*, panel B). In contrast, in serial mediation models, the *d*-link is from *M1* to *M2* (*d2* in *Fig. 4*, panel B) or from *M2* to *M1* (*d3* in *Fig. 4*, panel B). Although each specification of the *d*-link produces a theoretically distinct model, the models are statistically equivalent (Pieters 2017). Therefore, the different patterns of the *d*-link must be based on different theoretical assumptions and/or prior findings. The most appropriate model must be determined by theoretical reasoning and researchers must be aware that inferences about the links between the particular mediators or between the mediator and the dependent variable "are causally undetermined" (Pieters 2017, p. 697). In the example above, endorser expertise and trustworthiness are seen as two dimensions of the source credibility model (Ohanian 1990). However, researchers should be aware that prior research has found trustworthiness and expertise to have discriminant validity, but to be moderately correlated (Ohanian 1990). The critical issue is whether a causal chain between the mediators can

be justified theoretically; in the example, it would remain difficult to determine a direction of influence. Higher expertise could lead to higher trustworthiness and vice versa. Therefore, a parallel mediation model could be assumed to be appropriate.

If the researchers plan to include multiple mediators and the theoretical analysis leads to the assumption that indeed there is a causal link between the mediators (*d2* or *d3* in *Fig. 4*, panel B), a *serial* model would be more appropriate. Here, the effect runs either from *X* via *M1* to *M2* and then to *Y* or from *X* via *M2* to *M1* and then to *Y*. Let us assume that the researchers come to the theoretical conclusion that congruence (low vs. high) between the endorser and the product (*X*) would influence consumers' attributions as to why the celebrity endorses the product (e. g., the celebrity likes the product vs. the celebrity endorses the product simply because s/he is paid to do so) *(M1*), which in turn influences the perceived trustworthiness of the endorser (*M2*) and finally intention to purchase the brand (*Y*) (see *Fig. 4*, panel D). A possible formulation of a hypothesis representing these assumptions may be as follows:

H_{Med_ser}: *If there is high congruence between the celebrity endorser and the endorsed brand, then intention to purchase the brand will be more positive than in the case of low congruence. This effect is proposed to be mediated serially in the way that congruence impact consumers' attributions, which in turn impact perceived trustworthiness of the endorser, with purchase intention as the final consequence.*

Can I combine moderation and mediation?

Mediation analysis can be combined with the idea of moderation. The mediation model can be designed to include moderation of only the direct effect, of only the indirect effect or moderation of both the direct and the indirect effects (Hayes 2018). If the researchers are interested in whether a mediation effect ($X \rightarrow M \rightarrow Y$) functions differently in different contexts or for different people (i. e., in different conditions), moderated mediation can be applied (also referred to as conditional process analysis, Hayes 2018).

In a moderated mediation model, basically the same logic applies as in the simple moderation model: an indirect effect is the product of the effect of *X* on *M* and the effect of *M* on *Y*, controlling for *X*. The direct effect is the effect of *X* on *Y* controlling for *M*. But in a moderated mediation model, the indirect effect is contingent on the moderator. Therefore, "conditional indirect effects" (Hayes 2018, p. 393) for various values of the moderator can be estimated.

Reconsider the celebrity endorsement example where we thought about an effect of the congruency between celebrity endorser and the product (*X*) on purchase intention (*Y*) via perceived endorser trustworthiness (*M*). If we consider moderated mediation, we would ask whether there are boundary conditions for this indirect effect. A potential moderator *Z* could be persuasion knowledge (Friestad and Wright 1994). We would assume that the indirect effect (congruency → trustworthiness → purchase intention) would be stronger for consumers with low persuasion knowledge, whereas consumers with high persuasion knowledge would see through this persuasion tactic which would result in a weaker or even absent indirect effect (see *Fig. 5*, panel A). The hypothesis could be formulated as follows:

H_{ModMed}: *The indirect effect of celebrity endorser-product congruence on intention to purchase the brand via perceived trustworthiness of the endorser is stronger (weaker) in cases of low (high) persuasion knowledge of consumers.*

Researchers are strongly recommended to test for statistically significant differences between the conditional indirect effects (Pieters 2017). Of particular importance is that it is not enough to assume moderation of an indirect effect when one of the conditional effects has proved to be significant while the other has not. Instead, a formal test (e. g., "index of moderated mediation," Hayes 2018, p. 426) for the significant difference between the two (or more) conditional indirect effects is necessary (Pieters 2017).

The moderator can consider different parts of the mediation. Imagine the basic mediation model with one mediator *M*. Here, "first-stage moderation" and "second-stage moderation" (Pieters 2017, p. 694) can be distinguished. In first-stage moderation, the moderator influences the effect of the independent variable on the mediator (see *Fig. 5*, panel A). In second-stage moderation, the moderator influences the effect of the mediator on the dependent variable (see *Fig. 5*, panel B). The moderator can also exert influence on both stages (see *Fig. 5*, panel C). There are further models of moderated mediation (see Hayes 2018), including models that combine moderation and parallel or serial mediation.

3.2. Frequently asked questions about conducting experiments

Having formulated the hypotheses, researchers have to "translate" them into an experimental design. It is recommended that the reader considers works on the particular forms of experimental design (Field and Hole 2003; Shadish et al. 2002) or overviews on experimental designs that enjoy wide application in marketing research (Vargas et al. 2017), such as the basic random experiment, the randomized factorial design, or the randomized pretest-post-test design. In the following, we again turn to questions that may arise while making a decision regarding types of experimental designs and their practical arrangements.

What should I do if random assignment of participants to conditions is not possible?

Random assignment is a technique by which test participants "are assigned to receive the treatment or an alterna-

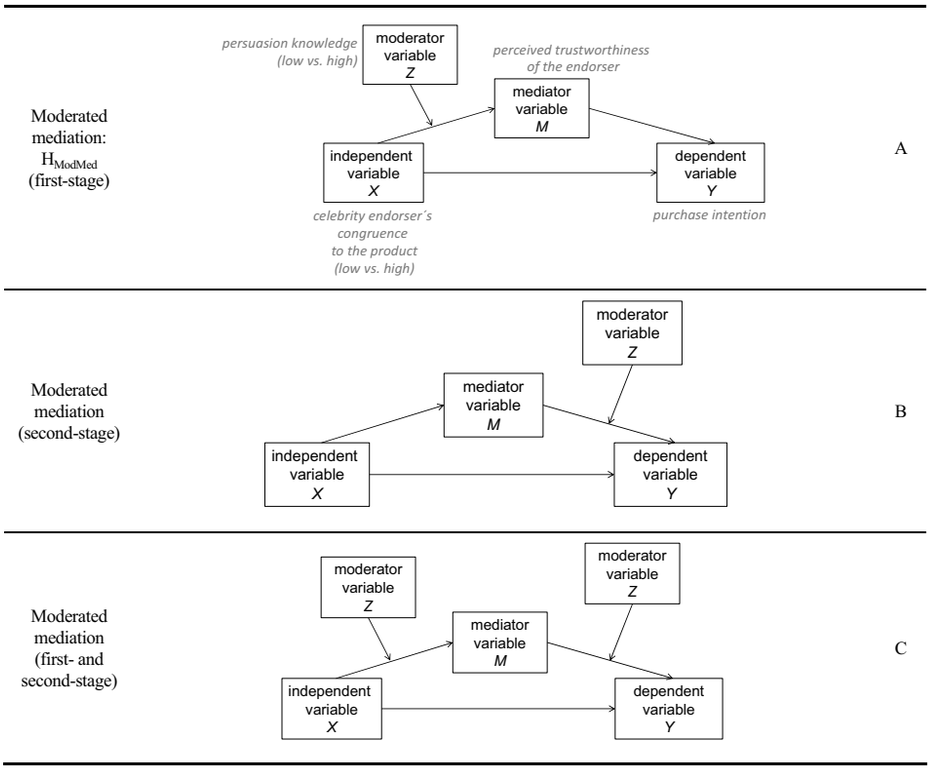

Fig. 5: Illustration of different forms of moderated mediation

tive condition by a random process such as the toss of a coin or a table of random numbers" (Shadish et al. 2002, p. 12). The reason for random assignment is to ensure that two or more groups of participants are similar to each other in terms of pre-treatment attributes. "Hence, any outcome differences that are observed between those groups at the end of a study are likely to be due to treatment, not to differences between the groups that already existed at the start of the study" (Shadish et al. 2002, p. 13).

Sometimes, random assignment is not possible or applicable. Imagine, for example, a field experiment using two different stores as settings for the manipulation of a marketing instrument: the tempo of music played is slow in one store and fast in the other. Customers in the stores should serve as test participants and their duration in the store is measured. However, when real customers are considered, researchers are not able to assign the customers to one of the two conditions (slow vs. fast music store), since the "treatment and control groups are intact, or already formed ... self-selected" (Kline 2009, p. 93). In cases like this, we touch on the differentiation of "randomized experiments" and "quasi-experiments." Using two real stores with real customers as experimental and control groups would result in a quasi-experimental design since customers self-select the store, with the result that there is no random assignment. Another issue related to the differentiation between randomized experiments and quasi-experiments comes into play when individual difference variables are examined. Individual difference variables such as gender, age, socio-economic variables or psychographic variables, such as differences in personality dimensions, can also be used as moderators. However, employing gender as a moderator variable implies that researchers are not able to "assign" participants randomly to either the male or the female condition because these are attributes inherent in the participants and cannot be manipulated.

In these situations, it is important to approach group equivalence in other pre-treatment attributes by means other than random assignment. One solution is to hold other variables constant. In the retail store example, the researchers need to think about characteristics that could be important to approximate group similarity. For example, the two stores should be similar in terms of product range, price range, socio-economic characteristics of customer base, etc. These and other relevant variables should be measured in order to check for non-significant differences between the customers of the particular stores (conditions). Other solutions are to select a stratified sample or to use matched samples (Geuens and De Pelsmacker 2017; Malhotra and Birks 2017). Kline (2009, p. 92) suggests that "with sufficient controls, a quasi-experimental design can be a powerful tool for evaluating causal hypotheses. This is why a well-controlled quasi-experimental design should not be viewed as the shabby, dirt-poor cousin of an experimental design, especially when it is impossible to use randomization."

Should I use a between-subjects or a within-subjects design?

In a between-subjects design, the participants are randomly assigned to the experimental conditions in a way that ensures that each participant is exposed to only one

experimental condition. In contrast, in a within-subjects design, each participant is exposed to multiple experimental conditions (Koschate-Fischer and Schandelmeier 2014). Reconsider the celebrity endorsement example from the beginning of this article. Erfgen et al. (2015) selected a between-subjects design. Each participant was shown either the advertisement presenting the celebrity endorser or the advertisement with the unknown endorser. In both conditions, brand recall was measured. A within-subjects design in which the participant would have been shown both advertisements would have made no sense in this case since the exposure to the brand in the first ad and the measurement of brand recall would influence recall of the brand in the second ad. This is referred to as testing effect or carry-over effect. Typically, testing effects result from taking a measure on the dependent variable more than once or before and after exposure to the treatment. Prior measures can affect later measures – for example, when respondents try to maintain consistency between pre- and post-test measures or when people become sensitized to a topic by a pre-measurement and pay more attention to it during the experiment than people who are not included in the experiment (Malhotra et al. 2017). Within-subjects designs are especially prone to testing effects. On the other hand, in within-subjects designs, fewer participants are required and they possess the strength of eliminating selection bias (Koschate-Fischer and Schandelmeier 2014). Selection bias describes a situation whereby treatment groups systematically differ with regard to characteristics related to the dependent variable before exposure to the treatment conditions. In the examples above, purchase intention might be influenced further by differences in the participants' income. When the effect is calculated "within subjects," such an influence is eliminated (Koschate-Fischer and Schandelmeier 2014).

While within-subjects designs suffer from testing effects and provide the benefit of eliminating a potential selection bias, the reverse is usually the case for between-subjects designs. Between-subjects designs reduce the possibility of participants seeing through the experimental design by measuring only once. However, in between-subjects designs, where different test participants are compared, selection bias may occur. Selection bias can be dealt with by random assignment, or, if random assignment is not possible, by the control of potentially relevant variables or by altering the potentially biasing variable to a constant (e. g., only considering participants belonging to a particular income group, if possible) in order to ensure no systematic differences between the groups.

Charness et al. (2012), Field and Hole (2003), Koschate-Fischer and Schandelmeier (2014), and Meyvis and Van Osselaer (2018) provide guidance for deciding between within-subjects and between-subjects designs.

Should I manipulate or measure variables?

Independent variables. In experiments, the independent variable typically is manipulated. Manipulation is a technical term, meaning that the independent variable has a set of varying levels and these levels are systematically changed *by the researchers* in order to assign participants to either an experimental group or a control group. Manipulation is used to achieve the criteria of temporal order that is necessary to derive causation. When respondents are exposed to the different levels of the independent variable which is followed by the measurement of the dependent variables, a temporal order of the independent and dependent variables is determined. Therefore, from the perspective of causal inference, manipulating predictor variables is preferable to measuring them.

How can researchers manipulate the independent variable? Usual ways are presenting written, verbal, or visual material to the participants by instructions and stimulus presentations in a written form (vignettes), video or computer (Cozby 2005). Reconsider the celebrity endorsement example: Erfgen et al. (2015) compared an advertisement containing a celebrity endorsement with an advertisement that pictured an unknown model. To manipulate the type of advertisement, they created two versions: one with a celebrity and one with an unknown endorser. In developing the advertising stimuli, the researchers had to select an appropriate celebrity endorser (e. g., Cindy Crawford, Heidi Klum, selected by Erfgen et al. 2015) and look for pictures of the celebrity, and also find an unknown endorser and select a similar picture in terms of position, smile, color of hair, etc. to that of the celebrity. The only difference between the endorsers should be that one is famous and one is not. To create the ad stimuli, researchers also had to select a product and a brand to be endorsed. Often, these selections are based on pretest data.

In the main study, the researchers would check for success of the manipulation (see section 3.3 "Which checks should I consider before testing the hypothesis?"). Here, they would check whether the celebrity endorser was actually known by the test participants, whereas the unknown endorser would actually not be known. Beside these "straightforward manipulations" (Cozby 2005, p. 167), there is the option of using "staged or event manipulations" (Cozby 2005, p. 169), which include simulating situations (e. g., a sales conversation or negotiation) or creating a particular psychological state in the test participants (e. g., customer dissatisfaction in order to analyze complaining behavior). Staged manipulations often are implemented by role-playing and employing confederates (who appear to be other participants in the experiment but are actually part of the manipulation; Cozby 2005) and include a higher extent of deception (see section 3.2 "Should I tell respondents what the study is about?").

Moderator variables. For the moderator variable, the same principle applies concerning manipulation as for the independent variable. Moderators can also be implemented by test stimuli. For example, in the Erfgen et al. (2015) study, congruence between endorser and product

was considered a potential moderator. The researchers had to select appropriate products and celebrity endorsers that fit vs. do not fit each other and expose participants to these conditions. The manipulation of congruence has to be checked within the main study.

In marketing experiments, moderator variables often relate to individual difference variables such as personality factors (e. g., extraversion, openness to innovation, self-monitoring), motivational factors (e. g., need for cognition, need for closure, regulatory focus), or relational variables (e. g., brand commitment, brand engagement). In these cases, manipulating the moderator is often difficult. Instead, a measurement of the moderator can be applied, and then the experiment becomes a quasi-experiment. However, in strengthening the claim of causal inference, researchers often use series of studies to avoid the criticism of potential correlation instead of causation. For example, in a study where the regulatory focus of consumers is considered a moderator, in a first study, the regulatory focus could be measured by using existing scales, whereas in a follow-up study, the regulatory focus could be manipulated by means of priming techniques.

Mediator variables. When thinking of mediation analysis, the first thought would be that mediators must be measured variables since they are also variables that are considered a reaction to the independent variables. This is what Spencer et al. (2005) refer to as *measurement-of-mediation design*. However, mediators can also be manipulated variables – they then become moderators in a *moderation-of-process design* (Spencer 2005). To be clear, it is not the aim of this section to mix up the understanding of moderation and mediation, although different authors have mentioned that these two are often confused (Baron and Kenny 1986; Spencer et al. 2005). To solve a potential confusion, according to Spencer et al. (2005, p. 847), researchers are recommended to "distinguish between theoretical and statistical understandings of mediation." A theoretical understanding of mediation would look for underlying processes for an effect of *X* on *Y*. Knowledge of the underlying process then refines the theory. However, from a statistical perspective, this mediating process can be tested in different ways. Among these, the Baron and Kenny approach (1986) and, more recently, the PROCESS macros available from Hayes (2013, 2018) are the approaches that most researchers are familiar with. Meanwhile, the PROCESS macro approach has been considered the standard approach (Geuens and De Pelsmacker 2017).

Another way to test statistically for an underlying process is to provide evidence that the effect of *X* on *Y* can be observed in some conditions *of a process variable* but not in others, or that the $X \rightarrow Y$ relationship is weaker vs. stronger in some conditions of the process variable than in others – a moderation-of-process design (Spencer et al. 2005). Why should researchers use this approach? Spencer et al. (2005), while also discussing the drawbacks of the approach, argue that "by manipulating both the independent variable and the mediating variable we can make strong inferences about the causal chain of events" (p. 846) because we utilize "the power of experiments to demonstrate causality" (p. 846). However, this approach can only be applied if the mediator variable can be easily manipulated, which is not always the case.

Dependent variables. The dependent variable is the outcome in the proposed causal chain. Therefore, dependent variables are typically measured or observed variables.

What should I take into consideration concerning the realism of the experiment?

The researchers have to decide on the experimental setting and whether the experiment should take place in a contrived or non-contrived environment. From a consumer behavior perspective, Morales et al. (2017, p. 472) differentiate between three types of experimental settings:

- *Field experiments*, where "participants do not know they are part of a research study *when* the manipulation is occurring and *when* they are engaging in real consumption behavior," which is observed or measured unobtrusively.
- *Realistic experiments in the field* which are "conducted outside the lab in actual consumption environments, but consumers are aware that they are taking part in a research study."
- *Lab experiments*, which are "conducted in controlled settings where participants are fully aware that they are part of a research study."

This differentiation is often connected with the differentiation of internal and external validity. Internal validity addresses the requirement that there should be no plausible explanations of the results *other* than those explanations that are considered by the independent variables. Laboratory experiments provide opportunities for higher internal validity since various factors that might influence the outcome can be controlled for (e. g., offers in stores, prices, advertising campaigns, crowding). However, laboratory experiments often appear less realistic, resulting in a lower external validity and therefore in limitations regarding the generalization of the results (but see discussion by Koschate-Fischer and Schandelmeier 2014). In contrast, field experiments appear to provide higher external validity (but see discussion by Gneezy 2017; Lynch 1999) and can be used to examine whether an effect really manifests under real-life conditions. However, in the field, it is often not possible to control for as many potential influences as in the laboratory, resulting in lower internal validity. Levitt and List (2009) offer discussion about different forms of field experiments with different advantages.

"Importantly, just as collecting data in the field does not necessarily make an experiment a field study, collecting data in the lab does not mean it has to be low in experimental realism or behavioral measures" (Morales et al.

Examples of contexts	Realism of the manipulation of the independent variable	Setting
Consumer behavior in sales conversations	Scenarios/vignettes describing fictive sales conversations	lab
	Use of voice recording or videos containing fictive sales conversations	lab
	Role playing to simulate sales conversations	lab
	Real sales conversations in the field	field
Product packaging evaluation	Pictures of products	lab
	Physical/real products	lab
	Real products on shelves/displays	lab
	Real products in simulated test stores	lab
	Real products in real stores in the field	field
Online consumer behavior	Pictures of webpages	lab
	Simulation of a website	lab
	Mock-up website	lab/field
	Real website	field
Negative publicity	Scenarios/vignettes describing fictive negative publicity on fictive brands	lab
	Scenarios/vignettes describing fictive negative publicity on real brands	lab
	Simulated newspaper pages describing fictive negative publicity on real brands	lab
	Real newspaper material describing actual negative publicity on real brands	lab/field

Tab. 4: Increasing realism of the manipulation of independent variables

2017, p. 472). Instead, Morales et al. (2017) consider the realism in experiments more generally and relate it to both the independent and the dependent variables. Concerning the independent variable, they recommend an experimental manipulation that involves real stimuli and entails a naturalistic setting. This is not to say that every experiment has to be in the field; however, the use of physical stimuli instead of pictures of the stimuli would already increase the realism in a laboratory experiment (see *Tab. 4*). Geuens and De Pelsmacker (2017) provide a discussion on whether to use stimuli that represent fictive or existing brand names, logos, or slogans.

Concerning the dependent variable, there is a call for more frequent consideration of behavioral measures in consumer behavior and marketing experiments instead of the commonly applied self-reported measures of behavioral intentions. Behavioral measures include choices, waiting time or time spent on the activity in question, writing recommendations, signing up, making a donation, as well as facial expression, eye movements/fixations or physiological responses (Morales et al. 2017; Woodside 2016).

Selection of an appropriate experimental setting depends on the research goal. Morales et al. (2017, p. 466) state: "[T]hough field-study data is often quite persuasive in convincing readers that an effect occurs outside the confines of a controlled lab environment, it rarely can provide any insight into the psychological underpinnings of a phenomenon and may not in fact be helpful for every paper." Gneezy (2017) recommends converging findings from laboratory and field experiments in series of studies. In one form, the researchers would start with a field experiment to show that an effect is present and relevant in practice, which would be followed by one or more laboratory studies to address the effect in more detail under controlled conditions. Another form works the other way around: one or more initial laboratory studies complemented by a field study (Gneezy 2017). Recent examples give an impression of field studies in marketing (see Meyer 2017). Gneezy (2017) also provides guidance regarding the practical steps of conducting a field study (for example, whether the experiment is conducted in collaboration with a partner).

Which items and scales should I use for measurement?

The experiment will include variables that need to be measured (the dependent variable, the mediating variables and all variables serving as controls and for manipulation checks and other checks). Within this issue of measurement, several decisions have to be made. Geuens and De Pelsmacker (2017) provide a comprehensive compilation of these sub-questions, including the following.

Why is extensive thinking about measurement techniques an issue at all? This question refers to the issues of content or construct validity that have to be ensured in order to adequately represent the constructs of interest. "Although one can certainly conduct statistical analyses on and with poor measures, the meaningfulness of any theory derived, theory test, or hypothesis examination becomes illusory with deficient measures" (Babin et al. 2016, p. 3135). Constructs should be defined clearly, and with high discriminant validity to other constructs. Items should reflect these conceptualizations. We have already discussed the particular relevance of discriminant validity in mediation models (Pieters 2017; Voorhees et al. 2016).

Which items should I use? Operationalization should embrace all aspects of the constructs. Many constructs have different dimensions that researchers should consider when selecting items (see examples in Geuens and De Pelsmacker 2017).

Should I adopt previously published scales? Researchers are generally recommended to use and cite previously developed, validated scales (taken from the relevant literature or from inventories, e. g., Bearden and Netemeyer 1999; Bruner 2015). However, even when validated scales are accessible, researchers should ascertain that the selected scale has high construct validity, particularly in the context of a different study (Babin et al. 2016; Geuens and De Pelsmacker 2017). In the case where the study addresses a construct that is new and has not been considered by previous scale developments or when important dimensions have been neglected thus far, researchers need to construct their own scales based on the methodological guidelines of scale development (Geuens and De Pelsmacker 2017; Bergkvist and Langner 2017).

Others have discussed the problem of *adopting versus adapting* measures taken from previous studies. Adopting means that "the current scale measure design is exactly as in its original form with no modification" (Ortinau 2011, p. 154). Adapting means that the original scale has been changed – either because researchers only use a portion of the original scale or because they make modifications in the item wording to fit the current context better (Ortinau 2011). Adapting is considered a limitation because it negates the previous findings on validity, reliability and dimensionality of the original scale (Babin et al. 2016; Ortinau 2011; Bergkvist and Langner 2017).

How many items should I use for each construct? This question relates to the decision between multi-item measurement and single-item measurement. Some discussion exists in the literature. While the classic approach recommends multiple measures, in some circumstances it is argued that single-item measures would be sufficient (Bergkvist and Rossiter 2007, 2009). Responding to the question of using single-item measurement, Diamantopoulos et al. (2012) provide guidance. Sarstedt et al. (2016a, 2016b) suggest caution in following the "trend" to consider single-item measures. In sum, Geuens and De Pelmacker (2017, p. 93) state, "[d]espite Bergkvist and Rossiter's (2007) argumentation, many researchers have their doubts, and it remains a challenge to convince reviewers and editors of the use of single-item measures for responses to advertising stimuli."

Should I also use reversed items? Reversed items are "items that need to be recoded to show a relation in the same direction with the underlying construct" (Geuens and De Pelsmacker 2017, p. 93). On the one hand, they can be used to enhance respondents' attentiveness, ensure that all aspects of the construct are grasped, counter respondents' tendency to agree, and help to avoid respondents seeing through the experimental research question. On the other hand, reversed items can confuse respondents and may threat the internal consistency of the items (see Geuens and De Pelsmacker 2017 for related references). Geuens and De Pelsmacker (2017) summarize how to formulate reversed items, which scale format should be used for them, and how they should be distributed throughout the questionnaire.

Which format of scales should I use? The scale format can bias responses owing to the response styles of participants. Decisions have to be made concerning the number of scale points, odd or even number of scale points, unipolar or bipolar scale format, scale numbering, and scale labels (see in detail Geuens and De Pelsmacker 2017; Bergkvist and Langner 2017).

In which order should I arrange the constructs in the questionnaire?

The measurement of one construct can impact responses to another construct. To avoid such order effects, Geuens and De Pelsmacker (2017, p. 88) recommend the following sequence: (1) introduction or briefing, (2) manipulation, (3) measurement of dependent variables, (4) measurement of items for quality control, (5) measurement of mediating variables and, if measured, of moderating variables, (6) measurement of potential confounds, (7) measurement of items for manipulation checks, (8) sociodemographic measures, (9) suspicion probe, and (10) debriefing.

Research on manipulation checks shows that certain effects can only be found if the manipulation check precedes the measurement of the dependent variables, which has to be considered a bias (Kühnen 2010). The early paper on manipulation checks by Perdue and Summers (1986) reviewed several problems that may arise from the timing of asking participants to agree or disagree with items aiming at manipulation checks and measurement of the dependent variable within the main study. They also discuss alternative solutions (e. g., manipulation check groups).

For the issue of using covariates, Meyvis and Van Osselaer (2018) discuss the sequence of measurement in order to achieve the criteria for including the covariate (see section 3.3 "When should I consider covariates?").

Must the independent variable be discrete and the dependent variable continuous?

In an experimental setting, the independent variables typically are manipulated. This leads to discrete variables (e. g., different versions of advertisements, different levels of fit between brand and brand extension product; manipulated mood: happy vs. sad; celebrity endorsers' attractiveness: low, moderate, high). Using a manipulated continuous independent variable would require as many experimental groups as there are levels of the variable (Vargas et al. 2017), which would seem to be ineffective, if not impossible. However, independent variables can be continuous if they are measured variables (e. g., personality traits; see section 3.2 "Should I manipulate or measure variables?").

It has to be noticed that in statistical mediation models, the independent variable is assumed to be a continuous variable, since mediation models are typically tested with regression-based approaches or structural equation modelling. However, with the help of dummy-coding, dichotomous independent variables can also be considered (Hayes 2018). More recently, approaches including multicategorical variables into regression-based moderation and/or mediations models have been discussed (Hayes 2018; Hayes and Preacher 2014; He et al. 2017).

The dependent variables can also be either discrete (e. g., choice between options) or continuous (or quasi-continuous such as rating scales). Regression-based approaches and analysis of variance that are typically used to analyze experimental data require continuous dependent variables. With the help of logistic regression it is also possible to include dependent variables that are not continuously measured.

What should I take into consideration concerning the use of random samples versus convenience samples?

There is a distinction between a sample that is drawn randomly from the population and convenience sampling, which does not select a random sample but uses test persons who are readily available. "Strictly speaking, no statistical inferences can be made without the selection of a random sample from a well-defined population to ensure that sample characteristics differ only by chance from the population characteristics" (Geuens and De Pelsmacker 2017, p. 86). However, experiments often rely on non-probabilistic samples; a review of the four leading advertising journals from 2008 to 2016 found probability sampling in only 8.2 % of the studies (Sarstedt et al. 2017).

Using non-probability samples is often considered acceptable until the study focuses on theory-testing (Calder et al. 1981; Geuens and De Pelsmacker 2017; Leary 2012). If the goal of the study is to describe how a population behaves, a random sample of that population is needed to approach this goal. However, often, the goal of marketing and consumer behavior studies is rather to test hypotheses regarding how certain variables relate to each other. "If the data are consistent with our hypotheses, they provide evidence in support of the theory regardless of the nature of our sample. Of course, we may wonder whether the results generalize to other samples, and we can assess the generalizability of the findings by trying to replicate the study on other samples of participants" (Leary 2012, p. 100). The discussion between the use of random samples or not, therefore, touches the differentiation between external and internal validity of the results (Espinosa and Ortinau 2016; Peterson and Merunka 2014). Babin et al. (2016, p. 3138) emphazise that in "tests with no intention for generalization ... [e]ffect sizes deserve greater attention relative to statistical significance."

From this perspective, the use of student samples (which is common in marketing experimental studies, see Espinosa and Ortinau 2016) can be interpreted. The critique (Espinosa and Ortinau 2016; James and Sonner 2001; Peterson 2001; also see discussion in Peterson and Merunka 2014) considers potential differences between students and "real people" in terms of age, income, education, lifestyle, experience with typical consumer products, and psychological characteristics (Ashraf and Merunka 2017). In a meta-analysis, Peterson (2001) found that in 19 % of studied relationships, variables related in different directions for students and non-students. This critique relates to the problem of using student samples for external validity. Therefore, Peterson (2001) demands that research based on students should be replicated with nonstudent participants which can be implemented by the use of series of studies with differing samples (in-built replications, Uncles and Kwok 2013). Furthermore, Peterson and Merunka (2014) tried to replicate effects found for a student sample with other student samples and obtained very mixed results. Therefore, they also demand greater consideration of replications by other researchers in order to analyze the reproducibility of results. In sum, Ashraf and Merunka (2017) provide detailed guidance concerning when it may be acceptable to use student samples.

Besides student samples, recent technological developments have given rise to the use of so-called crowdsourcing samples. Online crowdsourcing marketplaces are places where people are paid to complete tasks such as transcription, translation, photo tagging or participating in studies (Shank 2016). Web-based data collection platforms such as Amazon's Mechanical Turk (MTurk) enable fast and flexible recruitment of test participants at relatively low costs, making especially large-scale data collection more feasible. In addition, crowdsourcing samples are more heterogeneous than student samples and contain participants from different cultures (see overview by Goodman and Paolacci 2017). Regarding the use of MTurk participants in studies published in the *Journal of Consumer Research*, Goodman and Paolacci (2017) found an increase of studies using MTurkers, from 9 % of all studies in volume 39 (2012) to 43 % in volume 42 (2016).

Despite their popularity, crowdsourcing samples are not without challenges. From the perspective of external validity, a similar critique arises as for student samples: the question of whether these samples would represent the population. There is evidence that crowdsourcing samples deviate from the general (U.S.) population in important ways (see Shank 2016 for an overview). Furthermore, there is the threat of character misrepresentation that "occurs when a respondent deceitfully claims an identity, ownership, or behavior in order to qualify and be paid for completing a survey or behavioral research study" (Wessling et al. 2017, p. 211). Those who pretend in screening questions to have certain characteristics that are needed to complete the study (e. g., experience with certain products) may later give unstable answers that have no value for the researchers and cannot be used to

make generalizations about a population actually having these characteristics.

For crowdsourcing samples, there is concern for internal validity as well because the researchers lose control over the experiment, since the experiments do not take place in controlled settings and participants may be interrupted or distracted while working on the tasks (Shank 2016). While one important motivation for MTurkers is the financial incentive, there is the possibility that they are inattentive to instructions and provide poor-quality data (Chandler et al. 2014). In addition, owing to their growing experience with online tasks, "MTurk is a subject pool that *learns*, and its users often know more about social science research procedures than researchers may like" (Hauser and Schwarz 2016, p. 406), which might lead to different mental processes and reactions compared to "naïve" test participants (Goodman and Paolacci 2017; Hauser and Schwarz 2016).

Despite these concerns, the efficiency advantage of crowdsourcing pools has led to its popularity. Several researchers provide guidance into the use of crowdsourcing pools in order to address the concerns and ensure data quality (e. g., Goodman and Paolacci 2017; Wessling et al. 2017).

Should I tell respondents what the study is about?

There are ethical guidelines for treatment of participants (for experiments in marketing, see Geuens and De Pelsmacker 2017; Vargas et al. 2017; in general: Leary 2012), which include obtaining informed consent, that is to "explain to them what their participation entails and ask for their permission to be included in the study" (Geuens and De Pelsmacker 2017, p. 96). However, participants who know the true purpose of the experiment beforehand may not respond and behave as they would otherwise (hypothesis guessing, demand artifacts, Sawyer 1975; Allen 2004), leading to a decrease in internal validity. Experiments often involve some form of deception (Hertwig and Ortmann 2008), for instance by not informing participants about the existence of other experimental conditions or by using cover stories that conceal the true purpose of the experiment. Then, at the end of the experiment, when participants are no longer involved as participants, they must be informed of the true nature of the study by way of a debriefing (Vargas et al. 2017).

3.3. Frequently asked questions concerning data analysis

Which checks should I consider before testing the hypothesis? Can I remove cases if those checks indicate lower quality?

Different types of checks are needed or recommended before hypothesis testing. These include the manipulation check, confounding check, attention check, and quality checks. If data collection takes place over a longer period of time, time-invariance checks are recommended as well. The consideration of these checks also includes answering the question whether or not data from participants who did not pass a particular check should be excluded from further data analysis.

Manipulation check. The researchers have to check whether the manipulation of the independent variables (and the moderating variables, if included) was successful in terms of whether the stimuli used to represent the conditions of the independent variable are really able to mirror the underlying theoretical construct (Perdue and Summers 1986). For example, if the researchers would like to test whether advertisements containing a celebrity endorser (vs. an unknown model) result in better recall of the endorsed brand, they would select stimuli to manipulate celebrity endorsement (select a famous person as the spokesperson in the ad) and non-celebrity endorsement (select a non-famous person as the spokesperson for the other ad). The researchers would have to check whether the person selected as the celebrity endorser is really perceived as a celebrity by consumers. In the same way, they have to ensure that the person selected as the unknown person is not already known by consumers. Note that manipulation-check variables must not be used as independent or mediating variables. Their use is restricted to the manipulation check. In the later model, the manipulated variable (e. g., ad1 vs. ad2) is used.

There is a distinction between concrete, observable independent variables and unobservable independent variables (Perdue and Summers 1986). For observable variables, the manipulation can be confirmed objectively, perhaps because the studies include the manipulation of, for example, colors, man/woman as social stimuli, numbers (e. g., numbers of arguments or number of people depicted in an ad) (Geuens and De Pelsmacker 2017). Sometimes, attributes of the stimuli appear to be obvious, for example the tempo of music played in a store could be determined with objective measures and therefore manipulated objectively. However, it may be that different consumers *perceive* the tempo of music differently. Therefore, Geuens and De Pelsmacker (2017, p. 89) recommend using manipulation checks "whenever there is doubt about how 'obvious' manipulations are."

For unobservable variables, the manipulation can be perceived differently, as in the celebrity example above; therefore tests of these perceptions need to be conducted. It is recommended to check whether the manipulation of the stimuli is perceived as intended in a pretest as well as in the main study. A pretest is used to ensure that the stimuli are suitable for the main study. The main study contains the manipulation check to ensure that, for the sample that is used for the hypotheses tests, the manipulation was successful. However, to reduce the risk of hypothesis-guessing by participants, order effects of the questions used for the manipulation check should be considered (see recommendations by Perdue and Summers 1986, see section 3.2 "In which order should I arrange the constructs in the questionnaire?").

If two (or more) variables are manipulated, researchers have to ensure that each manipulation is successful on its own and that they do not interact in terms of the manipulation-check variables (Perdue and Summers 1986).

What should the researchers do if the manipulation check fails? If the stimuli selected are not perceived by the participants in the way the researchers had assumed, the intended manipulation has to be considered a fail (indicated by a non-significant difference between the conditions in the manipulation-check items). It is necessary to ascertain the reasons for this misperception, to find other, more suitable stimuli, then to pretest them carefully and conduct a new main study.

What should the researcher do if the manipulation check provides significant results but some of the participants fail to answer the questions in the way that was assumed before? This raises the question whether it is allowed to exclude these participants from the sample. It addresses the distinction between actual treatment and intention to treat (Meyvis and Van Osselaer 2018; Shadish et al. 2002). Because removing those participants where the treatment was unsuccessful (intention to treat) can easily result in confounding the data, Meyvis and Van Osselaer (2018) argue against mindlessly removing participants and offer criteria whereby excluding cases could be acceptable if strictly documented.

Confounding check. It should be ensured that the stimuli selected alter the manipulated construct only. This is to rule out "rival interpretations of what other constructs the manipulation might be varying" (Perdue and Summers 1986, p. 317). In the case of observable variables, "inadvertent confounding of the manipulations often can be avoided by maintaining *ceteris paribus* conditions across treatments" (Perdue and Summers 1986, p. 317). In the celebrity endorsement example, the product, brand, colors, slogans, size of ad person, position of the person in the ad, etc. should be the same between conditions. Confounding checks can also concern variables that are not directly observable but have been asked for. In the celebrity endorsement example, the ads should not differ in ways other than the famousness of the ad person, for instance by ensuring similar levels of attractiveness of the well-known and unknown endorsers that should be tested for in a pretest as well as in the main study.

What should the researchers do if the confounding tests fail? If the selected stimuli do not only differ in terms of the manipulation-check items but also in terms of the confounding-check items, potential effects of the independent variable on the dependent variable are referred to as "confounded." Consequently, it is not possible to draw valid conclusions from the experimental results. For example, consider again the celebrity endorsement example and assume that the two advertisements would differ not only in the perceived celebrity status of their particular endorser (successful manipulation check) but also in the perceived sympathy for the particular endorser (e. g., one of the endorsers is perceived as significantly more positive in terms of sympathy) indicating a failed confounding check. In this case, it would not be possible to infer a potential difference between the ads in brand recall (hypothesis) to the independent variable, because the confounding variable could serve as a second explanation for the effect. There is literature (Meyvis and Van Osselaer 2018; Yzerbyt et al. 2004) claiming that confounds can be "adjusted for" by the use of covariates; however, other voices consider interpretational problems (Field 2018; Miller and Chapman 2001; see section 3.3 "When should I consider covariates?"). The most certain and advisable way out would be to find out the reasons for the confounding, to select new stimuli based on these results, to pretest them carefully and to conduct a new study.

Quality checks. Several types of problems can arise when the respondents' work on the questionnaire is outside of the researchers' control. For example, in the case of online surveys, respondents may work on the questionnaire in a distracting environment, they may disrupt the answering procedure in order to work on other things and come back later, they may ignore the stimulus material, they may not read the questions carefully, or they may discuss questions with other people instead of giving an individual answer (Geuens and De Pelsmacker 2017). If the study is conducted online, the software usually measures the time participants spend on each page, providing the ability to identify respondents who took a break, provided answers faster than instructions were readable or needed an exceptional amount of time to answer the questions. Concerning attention to the stimulus material, control questions related to the contents of the stimuli can be included. Here, Meyvis and Van Osselaer (2018) again argue for paying attention to potential confounds when removing participants. Wessling et al. (2017) provide guidance for dealing with the problem of character misrepresentation among crowdsourcing sample participants.

Participants who do not read instructions carefully are assumed to be inattentive and to reduce the power of the experiment (Oppenheimer et al. 2009). Instructional manipulation checks (Oppenheimer et al. 2009), also referred to as "screeners" (Berinsky et al. 2014), are suggested as a way to detect participants who do not carefully read the instructions in a questionnaire. An instructional manipulation check (IMC) "consists of a question embedded within the experimental materials that is similar to the other questions in length and responses format ... However, unlike the other questions, the IMC asks participants to ignore the standard response format and instead provide a confirmation that they have read the instruction" (Oppenheimer et al. 2009, p. 867). This confirmation can include the request to answer in a predetermined way (e. g., "to show that you have read the instructions, please ignore the questions and select option A and B as your two answers"; see examples in Berinsky et al. 2014) or can include instructions that do not refer to the scale used (e. g., "in order to demonstrate that you

have read the instructions, please ignore the items below and simply click on the title at the top of this screen;" for examples, see Oppenheimer et al. 2009). The concern may arise that such screener questions "may signal to respondents that their answers are being monitored" (Berinsky et al. 2014, p. 744; Oppenheimer et al. 2009); however, as Berinsky et al. (2014) demonstrate, the use of screener questions does not induce social desirability bias. Berinsky et al. (2014) recommend detecting inattentiveness based on not only one but a set of different screener questions. However, MTurkers may become experienced with instructional manipulation checks *over time*, leading to higher attentiveness to the wordings or to different question interpretations than researchers had intended (Hauser and Schwarz 2016).

Several authors (e. g., Geuens and De Pelsmacker 2017, Meyvis and Van Osselaer 2018; Oppenheimer et al. 2009) recommend discarding participants who did not read the instructions carefully. However, as Berinsky et al. (2014) show, inattentiveness can be correlated with participants' characteristics. Therefore, an exclusion of those cases would skew the sample and induce bias. Instead, Berinsky et al. (2014) recommend reporting the experimental results for the full sample as well as for subsamples built on the basis of participants who passed the screener questions. If several screener questions are used, "researchers should present results stratified by attention ... [which] allows the readers to easily see how the results change as attention increases" (see Berinsky et al., 2014, p. 751, for an example of stratified reporting of results).

Time-invariance checks. The longer the time interval between observations, the higher the possibility that specific external events will confound an experiment (Malhotra et al. 2017). Such events may be economic conditions, brand crises, or social value discussions that suddenly become more present, e. g. in the media, and might increase participants' awareness of a topic of interest. The data should therefore be checked for differences in terms of time in order to rule out confounds by time ("history," Malhotra et al. 2017, p. 311). Short periods of experimentation, if possible, can be recommended as well.

When should I consider covariates?

An outcome variable is seldom influenced by only one independent variable. A variety of variables may exist that also have an impact. For instance, in a study examining the effect of front-of-pack food labels on consumers' intention to purchase the product, consumers' health consciousness may have an impact as well. Further variables like this – usually pre-treatment variables – that are not of theoretical interest in a particular experiment but are related to the dependent variable can be considered covariates.

Covariates can be included in the model. In *Fig. 6*, the variances of the three variables *X*, *Y* and the covariate *Cov* are represented by circles, with the overlap indicating shared variance (Miller and Chapman 2001). The logic behind covariate use is that the experimental effect can be partitioned into the explained variance (treatment effect, area 2 in *Fig.* 6) and the unexplained variance (because of factors that are not considered in the model – "noise", area 3). The covariate (as long as it is correlated with the dependent measure) represents an unconsidered factor whose impact is part of the unexplained variance (area 5). Including the covariate in the model reduces the unexplained variance (reduced area 3 in panel B compared to A) and increases test power (Field 2018).

Covariates should be defined *a priori*. But how can researchers know which covariates are likely to have an impact on the dependent variable before conducting the experiment? Thinking theoretically about potential influencing factors and taking into account prior studies on the effect can be helpful when considering the measurement or observation of covariates. Meyvis and Van Osselaer (2018, p. 1164ff) distinguish covariates that control for participants' individual differences in scale effects or response styles (e. g., extreme responding, tendency towards positive or negative answers), pre-existing indi-

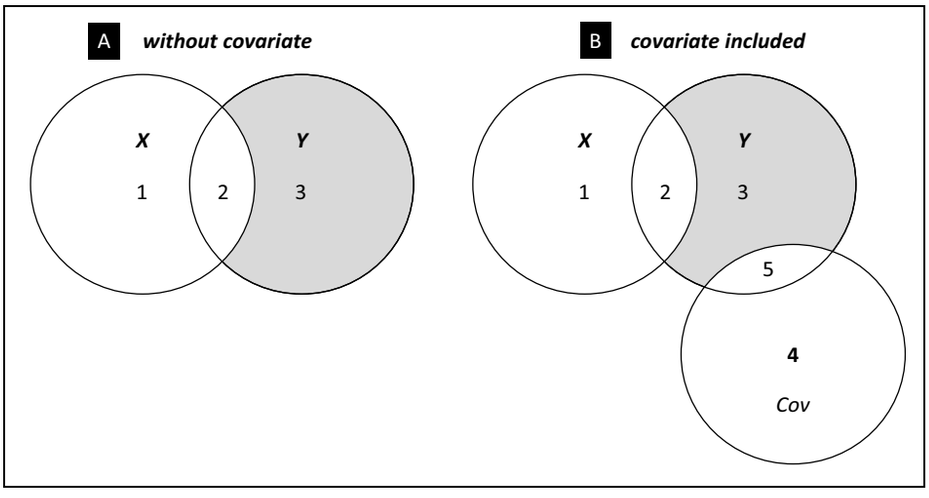

Fig. 6: Inclusion of a covariate

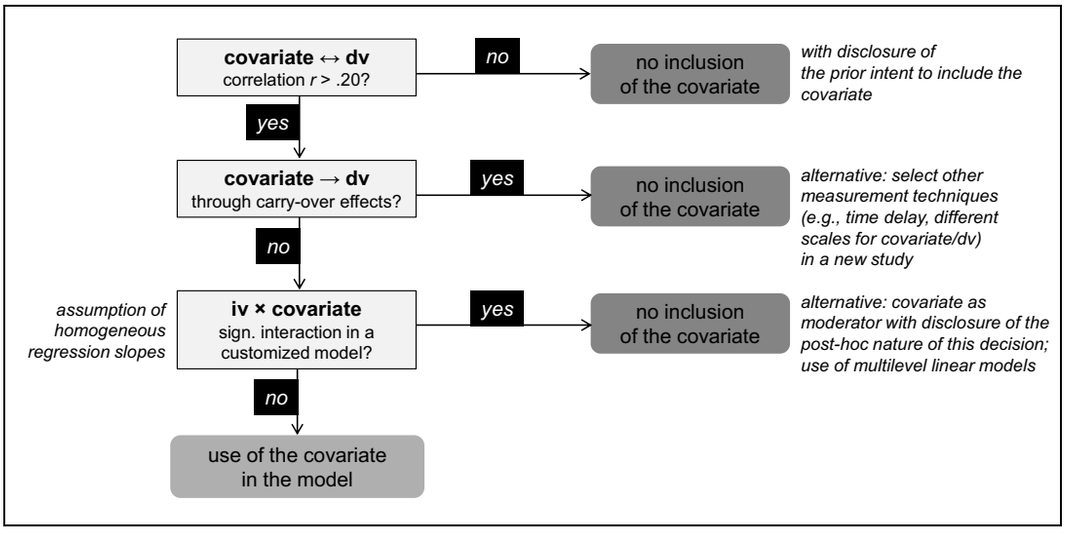

Fig. 7: Criteria for covariate inclusion

vidual differences in preferences between the participants (e. g., liking), stable individual differences in ability or cognitive resources (e. g., IQ, domain-specific knowledge or ability), individual differences in behavioral tendencies, demographics, or personality characteristics, and temporary individual differences (e. g., mood, tiredness). In series of studies that share the same dependent measures and stimuli, each study should consider the same covariate(s) in order to compare between studies (Meyvis and Van Osselaer 2018).

There are a number of criteria for the inclusion of a variable as a covariate into the model (see *Fig.* 7):

- *Correlation of covariate and Y.* The covariate should be correlated conceptually strongly with the dependent variable ($r > .2$) (Meyvis and Van Osselaer 2018). A planned covariate that was expected to be correlated with the dependent variable but eventually shows no substantial correlation should not be included in the model, but the prior intent should be disclosed (Meyvis and Van Osselaer 2018). When the covariate and *Y* are independent, the estimator of the effect of *X* on *Y* that ignores the covariate would perform better than the estimator of the effect that uses the irrelevant covariate information which also uses up degrees of freedom (Tabachnick and Fidell 2014).
- *No measurement effects of the covariate on Y.* Because covariates are usually considered pre-treatment variables that may exert influence on the dependent variables, the relationship between the covariate and the dependent variable should not originate from mere measurement effects or carry-over effects. To avoid carry-over effects, covariates should be measured with care. Meyvis and Van Osselaer (2018) offer recommendations for the formulation of the covariate measures, for the sequence of measurement in relation to the independent and dependent variables, and for the decision regarding similar or different scale formats for covariates and dependent variables. If there is a reason to assume carry-over effects in the experiment, the covariate should not be included.
- *Assumption of homogeneity of regression slopes*. The main reason for using covariates is their overall relationship with the dependent variable. At the same time, "we ignore the group to which a [participant] belongs. We assume that this relationship between covariate and outcome variable holds true for all groups of participants, which is known as the assumption of homogeneity of regression slopes" (Field 2018, p. 582). If the relationship between covariate and outcome is similar in all conditions of the independent variable *X*, then *X* and the covariate do not interact. To include a covariate in an analysis of covariance (ANCOVA), this assumption has to be tested using a customized model as demonstrated by Field (2018, p. 598). In the case of a significant interaction between covariate and independent variable (i. e., a violation of the assumption), the covariate should not be included in the model. Alternatively, it can be considered a moderating variable – taking the interaction into account in the model (Miller and Chapman 2001; Tabachnick and Fidell 2014; Meyvis and Van Osselaer 2018). This is an elegant solution to the problem; however, the empirically driven post-hoc nature of this decision must be disclosed (Meyvis and Van Osselaer 2018). An alternative is the use of a multilevel linear model (Tabachnick and Fidell 2014; Field 2018).

A further criterion that is discussed regarding the use of covariates is the question of whether the covariate needs to be independent from (uncorrelated with) the independent variable *X*. The ideal situation for covariate use is where the covariate is *uncorrelated with the independent variable*. This is the case when the randomization process has led to equivalence of samples across the experimental conditions (Geuens and De Pelsmacker 2017). In *Fig. 8* (panel A), zero correlation is represented by no overlap between *X* and *Cov*. Because there is no overlap,

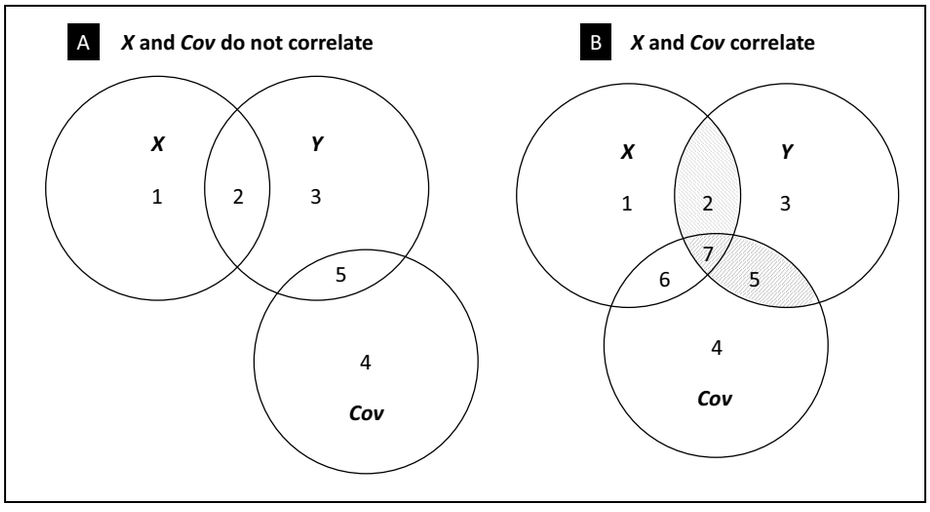

Fig. 8: Covariates uncorrelated and correlated with the independent variable

covariate inclusion is equivalent to a "noise-reduction" (Miller and Chapman 2001, p. 42) as described above.

In contrast, a situation may occur where the covariate is significantly *correlated with the independent variable* (*Fig. 8*, panel B, area 6 + area 7). In this case, the effect of the covariate overlaps the effect of the independent variable. Therefore, a potential treatment effect on *Y* is said to be "confounded" with the effect of the covariate. Because the conditions differ not only in terms of the independent variable but also in terms of the covariate, a difference between conditions in the dependent variable cannot unequivocally be inferred to the independent variable. In *Fig. 8* (panel B), the effect of *X* (area 2 + area 7) and the effect of *Cov* (area 5 + area 7) overlap in area 7.

What should the researchers consider in such a case? One point is that the inclusion of a covariate *as a statistical means* makes it possible to "remove" the influence of the covariate (on *X* and *Y*) from the model. However, a more important point is whether covariate use in the face of group differences on the covariate is *meaningful*. Although including covariates "to adjust for confounds" (Meyvis and Van Osselaer 2018, Yzerbyt et al. 2004) is a common practice, there are critics who question it from the perspective of interpretative issues (Miller and Chapman 2001). As Field (2018, p. 582) states: "the independence of the covariate and treatment ... is not a statistical requirement [for ANCOVA]", but dependence "creates an interpretational problem."

For approaching the statistical means of ANCOVA, the process of "adjusting for the effect of a covariate" can be illustrated by *Figure 8* (panel B). When conducting ANCOVA, the covariate is entered first into the model. This removes variance from *Y* that the covariate shares with *Y* (area 5 + area 7 in *Fig. 8*, panel B). This removal leaves portions of *Y* that are not related to the covariate (areas 2 and 3). Then the independent variable is entered into the model, followed by a calculation of the effect on adjusted scores and means of the dependent variable. This also results in the removal of some portion of the effect of *X* on *Y* (area 7) (Tabachnick and Fidell 2014).

However, from the perspective of interpretation, removal of the covariate effect is not considered to be appropriate in every situation. In approaching this interpretational problem, researchers are recommended to answer the following question based on their experimental design: *Can it be assumed that the overlap of X and the covariate has occurred purely by chance or has it occurred because of a substantial relationship between these variables?*

- *Overlap purely by chance:* This can be assumed when a random assignment of participants to the manipulated conditions of the independent variable was applied and the covariate has been measured before treatment. Although randomization is used to avoid differences in extraneous variables, such differences will occasionally arise by "chance rather than being meaningfully related to the group variable" (Miller and Chapman 2001, p. 40). In this case, ANCOVA would be appropriate. "The rationale for this view is that ANCOVA would only be removing noise variance ..., not anything substantive" (Miller and Chapman 2001, p. 45). After the removal, the remaining group variance would still be a valid representation of the construct (Miller and Chapman 2001).
- *Overlap based on substantial reasons: X causes Cov.* A case might arise where the manipulated independent variable has a causal impact on the covariate when the covariate is measured after treatment. For example, consider a test of different front-of-pack food labels on consumers' intention to purchase a product. When consumers' health consciousness is considered a covariate and measured after treatment, the different food labels could be assumed to have a causal impact on health consciousness. Therefore, "the manipulation of the independent variable should not cause differences in the level of the covariate" (Meyvis and Van Osselaer 2018, p. 1165). If the anal-

ysis reveals the possibility of a causal influence of the independent variable on the covariate, the covariate should not be included in the model. Instead, it can be considered a mediator (Meyvis and Van Osselaer 2018; Miller and Chapman 2001; Yzerbyt et al. 2004); however, the empirically driven nature of this decision should be disclosed. If the researchers can make a strong case that causal impact makes less sense, the covariate could be considered in the model. The best way would be to avoid causation completely by measuring the covariate before participants are randomly assigned to the experimental conditions. However, sometimes researchers refrain from placing the covariates at the beginning of the questionnaire because they fear hypothesis-guessing by the participants.

- *Overlap based on substantial reasons: X is systematically related to a defining characteristic of the groups.* This becomes relevant in non-randomized experiments where preexisting groups are studied (e. g., age groups, customer groups). It may then occur that these groups differ in certain pretreatment variables. These "observed pretreatment differences may reflect some meaningful, substantive differences that are attributable to group membership" (Miller and Chapman 2001, p. 40). For example, imagine that researchers would analyze differences between SINUS milieu groups (e. g., traditional milieu, hedonistic milieu, established milieu) in attitudes to a premium brand, and income was planned to be considered a covariate. However, income would be systematically related to the defining characteristic of the groups (e. g., with certain milieus having more or less income than others). Including income as a covariate in the analysis would result in the removal of shared variance between the covariate and the independent variable which could "corrupt the grouping variable itself" (Miller and Chapman 2001, p. 44), "leaving an undercharacterized, vestigial [remaining variance of the independent variable] with an uncertain relationship to the construct that [*X*] represented" (Miller and Chapman 2001, p. 45). The effect that remains after the removal of the covariate effect may be difficult to interpret. Therefore, where *X* and the covariate are conceptually closely related (resulting in a larger part of shared variance that would be removed from *X* during ANCOVA), the inclusion of the covariate cannot be recommended.

I have considered a moderator variable that has been measured as a continuous variable. Should I calculate categories for this variable in order to use ANOVA?

One-factorial ANOVA is used when the model includes only one binary or categorical independent variable. Two-factorial ANOVA or factorial ANOVA on higher levels is employed if the independent variable as well as the moderator variable(s) are binary or categorical variables. Regression-based approaches are used if the independent variable and/or the moderator are measured with continuous scales. Imagine, for example, a study that considers product involvement as a moderator. Researchers could manipulate product involvement (low vs. high), resulting in a discrete variable, and calculate the model using ANOVA. They could also measure product involvement as a continuous variable and use regression-based approaches (however, see section 3.2 "Should I manipulate or measure variables?").

Continuous independent variables can possibly be transformed into binary variables, a technique referred to as "dichotomizing." For example, researchers could split their sample using a cutting point such as the median (or the mean). This would result in two groups that would differentiate participants below vs. above the median of product involvement.

There is an ongoing discussion on whether such median splits are appropriate. In earlier articles, "dichotomizing" a continuous variable into two groups in order to be able to use ANOVA was often found. However, several authors (e. g., MacCallum et al. 2002) have warned that dichotomization of a continuous variable can be accompanied by serious problems such as spurious main or interaction effects (i. e., they are statistically significant but artificial results; Kline 2009, p. 50) or reduced statistical power (Irwin and McClelland 2003). As an alternative, in the case of continuous moderator variables in experimental designs, regression-based approaches are recommended (Fitzsimons 2008; Spiller et al. 2013). The researchers, then, would regress the dependent variable on the manipulated independent variable, the continuous moderator variable, and their interaction. The regression analysis provides the slopes and their significance for different values of the continuous moderator variable, typically for the mean of the moderator and values plus/minus one standard deviation from the mean (but see discussion by Spiller et al. 2013). The significant difference in the regression slopes is reflected in a significant interaction effect. With the clear address "death to dichotomizing" (Fitzsimons 2008, p. 5) and the emergence of SPSS macros which made regression-based moderation analyses more comfortable (Hayes 2013, 2018), dichotomizing no longer seemed to be an issue because "current thinking suggests that median splits will always produce inferior analytic conclusions" (Iacobucci et al. 2015b, p. 653), which affects the review process inasmuch as "a researcher who submits a paper that includes a median split is almost certain to provoke the ire of the review team" (Iacobucci et al. 2015b, p. 658). However, more recently, Iacobucci et al. (2015b) have argued that the issue of dichotomizing is much more nuanced than the critical articles suggest. They argue that dichotomizing a continuous variable is problematic in the case of multicollinearity between predictor variables. In the case of correlated predictors, therefore, researchers are recommended to use continuous scores, and median splits should not be employed. In the case that the variables can be confirmed to be uncorrelated, median splits are acceptable if the research interest lies in group differences (individuals who

score low vs. high on a construct). However, if variables are uncorrelated but the research interest lies in examining individual differences, the continuity of measured scales would be better to represent the construct. Then median split should not be used (see the matrix by Iacobucci et al. 2015b, p. 662).

These recommendations have provoked intensive discussion (Rucker et al. 2015; McClelland et al. 2015; Iacobucci et al. 2015c). A Google Scholar search in January 2018 for the citation of the Iacobucci et al. (2015b) recommendations yielded 75 hits, including several articles that now justify median splits in their empirical studies. However, it can be recommended to consider the available alternatives to dichotomization (see Spiller et al. 2013 for a tutorial).

Why should I include an effect size calculation?

An important issue is the differentiation between statistically significant findings ("Does the manipulation of, for example, a marketing instrument affect the dependent variable in a statistical sense?", e. g., $p < .05$?) and scientifically significant findings ("How effective is the tested instrument?"). This issue is considered by effect size indicators (e. g., eta squared, Cohen's *d*). Tests on statistical significance conflate sample size and effect size (Kline 2013). A relatively small effect can be of statistical significance when the sample size is large enough (has appropriate statistical power). Hayes (2018) and Lachowicz et al. (in press) provide reviews on various types of effect size indicators. They and others (e. g., Preacher and Kelley 2011; Prentice and Miller 1992) caution against mixing up effect size and importance (i. e., "small" effects can nevertheless be important) and stress that the study context has to be considered when evaluating effect sizes (Tabachnick and Fidell 2014).

Which parameters of the findings should be reported?

Several sources have provided guidance for the structure of reporting methodology of research studies (APA 2010) as well as for the reporting of findings of multivariate analysis methods (Tabachnick and Fidell 2014) or of experiments (e. g., Field and Hole 2003). Ortinau (2011) gives recommendations about what aspects the written work should contain. Bakker and Wicherts (2011) list types of reporting errors (e. g., incomplete statistics, rounding errors, inexactly reported *p*-values) that frequently arise in statistical sections of articles; this list can be used to raise awareness of which parameters and findings should be reported in which way. Depending on the particular data analysis method used, specific discussions have to be considered. For example, Peterson and Umesh (2018) discuss which parameters should be provided when using ANOVA; Pieters (2017) gives a set of recommendations on reporting mediation analysis results. Moreover, the usefulness of descriptive data for the variables under research is emphasized: "Readers may take more from these data than from statistically massaged results" (Babin et al. 2016, p. 3137). Of utmost importance is a full disclosure of how many participants have been excluded from the analyses and for what reasons (checks, outliers) and which covariates were planned a priori to be included in the analysis but were not considered afterwards for which reasons (Babin et al. 2016; Bakker and Wicherts 2014; Meyvis and Van Osselaer 2018).

Another issue related to reporting refers to the efforts researchers and journals undertake in generalizing research findings by study replications or meta-analyses. Lehmann and Bengart (2016) as well as Woodside (2016) call for reporting all information on the procedure and all findings that would be needed by other researchers to conduct study replications or meta-analyses. An oftenmentioned limitation that the authors of meta-analyses have to concede is that not all studies that are identified as relevant during the literature search process could be considered for the meta-analysis because critical parameters have not been reported by the original researchers.

A further problem that critical voices have identified is that studies with non-significant results may appear in a series of studies, but are not reported as a part of the series. In this case, misunderstandings on the actual robustness of the effects may arise in the academic community "because the published literature likely overstates those relationships" (Babin et al. 2016, p. 3137).

3.4. Frequently asked questions concerning interpretation of the findings

Calculating and reporting the experimental findings is not enough. The researchers are supposed to interpret their findings and set them in the context of previous research. This enables them to refine theory and to contribute to existing knowledge. Furthermore, for stressing the practical relevance of experimental research, implications for practice might be offered (Bartunek and Rynes 2010).

What should I do if the effect that I proposed turns out to be not what I expected?

Statistically non-significant results are not an exception. It is important that the researchers elaborate on the reasons for their non-significant results. Peterson and Umesh (2018) provide guidance for this elaboration process. Non-significant results are often attributed to power problems (e. g., small sample sizes). However, in some circumstances, very small *F*-statistics occur. Then, these very small *F*-statistics imply that the denominator of the *F*-ratio (the "error") is so large that even by chance one cannot expect such a large error relative to the treatment effect. This is referred to as a "statistically significant insignificant result" (Peterson and Umesh 2018, p. 82). In such cases, the non-significance should not be attributed to power problems or other smaller problems but should raise "a 'red flag' that suggests potential problems with the theory underlying the experiment and/or the design or implementation of the experiment itself" (Peterson

and Umesh 2018, p. 83). Causes of experimental failure can be design-related, related to violations of assumptions of the statistical models chosen, or research subject-related (Peterson and Umesh 2018). We have discussed several of these issues in this article. However, researchers who have carefully developed and conducted the experiment may be allowed to infer that such a very small *F*-statistic "would more likely seem to be prima facie evidence of a theoretical failure as opposed to an experimental failure" (Peterson and Umesh 2018, p. 85). Although authors state that such non-significant results are valuable (e. g. Babin et al. 2016) and are in line with Popper's (1959) notion of falsifiability of theories (Peterson and Umesh 2018), studies that fail to prove a theory are seldom published (Armstrong 2003; Peterson and Umesh 2018).

Sometimes the findings are statistically significant but show a different direction of effects from that which was expected. Given a comprehensive theoretical reasoning, these results are surprising because they indicate that the process assumed might not work as expected, but other processes might be responsible for the effects found. A result that shows the opposite direction than expected can also be regarded as an indicator of a hidden moderator. From a methodological perspective, researchers then can think about the differences in situational factors and context in their study in comparison to other studies with other results previously found. More generally, they can try to make sense of the surprising findings (but see the discussion by Field 2018, in terms of two-tailed vs. one-tailed tests). Potentially, a completely different theory must be used to explain these results. Eventually, surprising results can enable the researchers to elaborate on new hypotheses or even on new theories (Babin et al. 2016).

However, as Armstrong (2003) illustrates, controversial findings that differ from current beliefs or practice are not easy to publish. There is a call in the literature to consider unexpected results more openly (Armstrong 2003; Babin et al. 2016). Induction where researchers draw inferences from datasets and their results is a useful way of theory building. However, it is of utmost importance that the test of the new hypotheses is conducted with new data sets containing other participants. Presenting post-hoc findings as a priori hypotheses is one of the questionable research practices that Banks et al. (2016) identified among researchers.

The majority of journal articles seem to find support for the hypotheses (Babin et al. 2016). Reasons may be that researchers test for very safe hypotheses or that the system of publishing rewards the finding of hypothesis-confirming results (Banks et al. 2016). Consequences are that researchers may "cherry-pick" and report only those results that show support. This practice may lead to misunderstanding and biased impressions of the real occurrence of an effect in the scientific community (Babin et al. 2016). Moreover, it may lead to frustration on behalf of other researchers who have to face non-significant results but are not aware of this publication bias.

The misguided pursuit of hypotheses-confirming results also includes questionable research practices such as neglecting an experimentally manipulated variable in the process of data analysis or reinterpreting a priori control variables ex-post as experimental factors without making clear the nature of a post-hoc decision. The latter may also be a reason why ex-post dichotomization of variables evokes skepticism among reviewers. For instance, if involvement as an experimental factor comes from a median split of a continuous measure, one could wonder why the researchers did not manipulate this factor systematically from the outset, for instance by means of a priming task. To sum up, if initially expected effects turn out to be non-significant or show unexpected signs, researchers are expected to refrain from "going on a fishing expedition."

What should I take into consideration concerning the generalizability of the results?

Generalizability is an issue of external validity. "External validity examines whether or not an observed causal relationship should be generalized to and across different measures, persons, settings, and times" (Calder et al. 1982, p. 240). Uncles and Kwok (2013) provide guidance for so-called "in-built" replications that refer to replications conducted by the researchers of the initial study themselves by varying the studies' attributes of content (e. g., varying product categories), place (sales areas, nations, cultures), and time. Morales et al. (2017) discuss the possibility of generalizing from non-behavioral measure to behavioral measures.

In addition, there is a loud call for external replications to be conducted by other researchers. To address these concerns, some journals have reserved special space for replication studies (e. g., the replication corner in the *International Journal of Research in Marketing*). Different types of replications can be distinguished: direct replications (using exactly the same procedure as the original), conceptual replications and replications with extensions. Conceptual replications study the same concept-to-concept relationships as the original article but operationalization may differ from the original (other segments of participants, other procedures). Replications with extensions study the original while considering an additional moderator (Lynch et al. 2015). Conceptual replications are considered superior to direct replications (Crandall and Sherman 2016; Lynch et al. 2015).

4. Conclusions

Experimental research relies on assumptions about cause and effect. Experimental researchers need to know how to manipulate independent variables and moderator variables and study their effects on the dependent variable, while controlling for other potential explanations of the effects; these are the tools that they must master. In the present article, we have provided an overview of some general and some specific decisions that experimental researchers must consider when preparing and conducting

experiments, analyzing experimental data and interpreting results. We have linked our description of questions to recent discussions in the literature and provided the reader with further references where relevant.

Regarding the theoretical preparation of experiments, we have discussed different forms of hypotheses (basic form, moderation, mediation, integrated moderation and mediation), their interpretation and points to consider when deriving them. Regarding the conducting of experiments, we have discussed decisions about the design of variables (manipulation or measurement, realism), samples and ethical issues. Regarding data analysis, we have discussed the necessary checks, covariation, dichotomization, effect size and issues relating to the reporting of results. Finally, regarding the interpretation of results, we have considered the non-significance of findings and the generalization of results.

Of course, novice experimenters may have further questions. For example, they may want to know about the use of incentives and their impact on participants' willingness to participate and their responses (Espinosa and Ortinau 2016; Koschate-Fischer and Schandelmeier 2014). They may want to know about the use of mean-centering (Dawson 2014; Hayes 2018; Iacobucci et al. 2015a), or the use of one-tailed or two-tailed testing for directional hypotheses (Cho and Abe 2013; Field 2018), the use of structural-equation modeling for the analysis of experimental data or mediation models (Bagozzi and Yi 1989; Hayes et al. 2017; Iacobucci et al. 2007) – the list goes on.

Our review of literature found an astonishingly large number of methodological articles on fundamental issues in experimentation published in recent years. Evidently, the replication crisis in social psychology and related areas has led to demand for more guidance by experienced researchers (Meyvis and Van Osselaer 2018). Novice researchers are advised to avoid merely citing authors of other experimental studies who use appealing procedures ("they did it that way, therefore so do I") and instead to provide arguments in favor of their procedures, drawing attention to recent developments in experimentation. Several journals feature regular tutorials or special issues on methodological problems. Recent discussions in the literature on the fundamentals of experimental contributions (East 2016) and on issues of execution, analysis and interpretation indicate that much remains to be said on the topic of experimental research.

Our investigation of the current discussion in the literature reveals four particularly interesting developments that we believe it will be worth monitoring going forward. They are as follows:

- *Using competing hypotheses.* As early as 2001, Armstrong et al. stressed that overuse of the "dominant hypothesis" approach could lead to biased results. Despite this, even today very few articles use the "competing hypotheses" approach.
- *Using measurement-of-mediation and moderation-of-process.* A citation analysis by Demming et al. (2017) reveals that statistical mediation analysis has increased in the last decade. From a theoretical perspective this is understandable, as analyzing the underlying processes of an $X \rightarrow Y$ effect clarifies "what would otherwise remain a black box in terms of why a manipulated stimulus predicts an outcome" (Geuens and De Pelsmacker 2017, p. 89f). However, the original goal of experimentation – analyzing causal chains – may be put at risk, as the mediator and dependent variable are simply correlated, while the causality between them remains undetermined (Pieters 2017) or must be argued on a theoretical basis. As a methodological solution, some researchers have proposed moderation-of-process designs (Spencer et al. 2005). These designs go back to the original idea of experimentation, that of proving cause-and-effect relationships. Spencer et al. (2005) already considered the statistical mediation analysis overused back in 2005 (when the Baron and Kenny (1986) approach was the standard), and the perception that this is the case may be even stronger today now that the PROCESS macro (Hayes 2013, 2018) has made mediation analysis easier and even more widespread (Demming et al. 2017). It will be interesting to monitor whether the suggested use of moderation-of-process designs by Spencer et al. (2015) and the critical points raised by Pieters (2017) on mediation analysis and its use for analyzing causal chains, which have also been addressed in applied disciplines such as advertising research (Geuens and De Pelmacker 2017), will encourage researchers to use experimental designs more often when analyzing underlying processes. This would also mean that novice researchers would have to distinguish more clearly between the theoretical and statistical meanings of moderation and mediation.
- *Crowdsourcing samples.* Further research is necessary to determine the appropriateness of crowdsourcing samples. While some articles suggest that the potential problems are manageable by means of quality checks, others warn against their use. Some critics consider the possible side-effects of this increasingly popular sampling technique. For example, Pham (2013, p. 420) states that "there is a real danger of the low data collection costs associated with MTurks gradually shifting our research agendas toward studies that can be done using MTurks [...] as opposed to studies that should be conducted to advance our field."
- *Reporting non-significant findings.* Babin et al. (2016, p. 3137) state that the "bottom-line is that nonresults are equally important to results in understanding the real world." However, researchers may be afraid of putting themselves at a disadvantage by reporting non-significant results and therefore not fully disclose those results. Banks et al. (2016) consider withholding results and selective reporting of results a common, yet questionable research practice. The problem ap-

pears to be motivated by the current system of "publication practices that implicitly reward the finding of significant results that confirm study hypotheses" (Banks et al. 2016, p. 329). Journals and reviewers should therefore strongly encourage researchers to disclose non-significant or surprising results. It might be useful to establish submission formats such as the "hybrid registered reports submission" (Banks et al. 2016), whereby articles are reviewed on the basis of their conceptual and methodological parts, irrespective of whether the findings – submitted at a later stage – are significant or not.

Notes

[1] In this article, we focus on independent variables with two conditions that allow comparisons between two manipulated groups. In this situation, the effect can be assumed to be positive (compared to the first group, the second group shows higher values on the dependent variable) or negative (the second group shows lower values). In addition, there is the possibility of considering more than two groups.

[2] The label "two-way interaction" ("three-way interaction") describes that two (three) independent variables are considered to interact. There is no label "one-way interaction" because it takes at least two variables to interact. However, a "one-way ANOVA" is used to test for differences between the levels of one independent variable.

[3] Voorhees et al. (2016) recommend experimental studies using multi-item scales to test for discriminant validity using the HTMT technique (Henseler et al. 2015) or the AVE-SV technique (Fornell and Larcker 1981).

[4] Causal directionality refers to the question whether it can be assumed that a predictor precedes an outcome in a causal relationship ($X \rightarrow Y$, $M \rightarrow Y$), excluding the possibility of the reversed direction ($Y \rightarrow X$, $Y \rightarrow M$) or a mutual relationship ($X \leftrightarrow Y$, $M \leftrightarrow Y$). Note that causal directionality is a concept other than the directionality or non-directionality of a research hypothesis. In undirectional hypotheses it is undetermined whether the (causal) effect is a positive or negative one, while in directional hypotheses the positivity or negativity of the effect is determined.

[5] Although parallel mediators can be correlated, researchers should take collinearity into account. Preacher and Hayes (2008, p. 887) recommend to "select mediators that represent unique constructs with as little conceptual overlap as possible. Following this strategy will minimize collinearity."

References

Aguinis, H., Edwards, J. R., & Bradley, K. J. (2017). Improving Our Understanding of Moderation and Mediation in Strategic Management Research. *Organizational Research Methods*, 20(4), 665–685.

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple Regression: Testing and Interpreting Interactions*, Thousand Oaks, CA: Sage.

Allen, C. T. (2004). A theory-based approach for improving demand artifact assessment in advertising experiments. *Journal of Advertising*, 33(2), 63–73.

Andersson, U., Cuervo-Cazurra, A., & Nielsen, B. B. (2014). From the Editors: Explaining Interaction Effects within and across Levels of Analysis. *Journal of International Business Studies*, 45(9), 1063–1071.

APA – American Psychological Association (2010). *Publication Manual of the American Psychological Association*, 6th ed., Washington, DC: American Psychological Association.

Armstrong, S. J. (2003). Discovery and Communication of Important Marketing Findings. *Journal of Business Research*, 56(1), 69–84.

Armstrong, S. J., Brodie, R. J., & Parsons, A. G. (2001). Hypotheses in Marketing Science: Literature Review and Publication Audit. *Marketing Letters*, 12(2), 171–187.

Ashraf, R., & Merunka, D. (2017). The Use and Misuse of Student Samples: An Empirical Investigation of European Marketing Research. *Journal of Consumer Behavior*, 16(4), 295–308.

Babin, B. J., Griffin, M., & Hair, J. F. (2016). Heresies and Sacred Cows in Scholarly Marketing Publications. *Journal of Business Research*, 69(8), 3133–3138.

Bagchi, R., Block, L., Hamilton, R. W., & Ozanne, J. L. (2017). A Field Guide for the Review Process: Writing and Responding to Peer Reviews. *Journal of Consumer Research*, 43(5), 860–872.

Bagozzi, R. P., & Yi, Y. (1989). On the Use of Structural Equation Models in Experimental Designs. *Journal of Marketing Research*, 26(3), 271–284.

Bakker, M., & Wicherts, J. M. (2011). The (Mis)Reporting of Statistical Results in Psychology Journals. *Behavioral Research*, 43(3), 666–678.

Bakker, M., & Wicherts, J. M. (2014). Outlier Removal, Sum Scores, and the Inflation of the Type I Error Rate in Independent Samples t Tests: The Power of Alternatives and Recommendations. *Psychological Methods*, 19(3), 409–427.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly. *Journal of Business Psychology*, 31(3), 323–338.

Baron, R. M., & Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research. Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.

Bartunek, J. M., & Rynes, S. L. (2010). The Construction and Contributions of "Implications for Practice": What's in Them and What Might They Offer? *Academy of Management Learning and Education*, 9(1), 100–117.

Bearden, W., & Netemeyer, R. (1999). *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*. Thousand Oaks, CA: Sage.

Bergkvist, L., & Langner, T. (2017). Construct Measurement in Advertising Research. *Journal of Advertising*, 46(1), 129–140.

Bergkvist, L., & Rossiter, J. R. (2007). The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs. *Journal of Marketing Research*, 44(2), 175–184.

Bergkvist, L., & Rossiter, J. R. (2009). Tailor-Made Single-Item Measures of Doubly Concrete Constructs. *International Journal of Advertising*, 28(4), 607–621.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science*, 58(3), 739–753.

Bouwman, R., & Grimmelikhuijsen, S. (2016). Experimental Public Administration from 1992 to 2014. *International Journal of Public Sector Management*, 29(2), 110–131.

Brown, J. R., & Dant, R. P. (2008). On What Makes a Significant Contribution to the Retailing Literature. *Journal of Retailing*, 84(2), 131–135.

Bruner, G. C. (2015). *Marketing Scales Handbook. Multi-Item Measures for Consumer Insight Research*. Fort Worth, Tx: GCBII Productions.

Calder, B. J., Philipps, L. W., & Tybout, A. M. (1981). Designing Research for Application. *Journal of Consumer Research*, 8(2), 197–207.

Calder, B. J., Philipps, L. W., & Tybout, A. M. (1982). The Concept of External Validity. *Journal of Consumer Research*, 9(3), 240–244.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and

Solutions for Behavioral Researchers. *Behavior Research Methods*, 46(1), 112–130.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1–8.

Chennamaneni, P. R., Echambadi, R., Hess, J. D., & Syam, N. (2016). Diagnosing Harmful Collinearity in Moderated Regressions: A Roadmap. *International Journal of Research in Marketing*, 33(1), 172–182.

Cho, H. C., & Abe, S. (2013). Is Two-Tailed Testing for Directional Research Hypotheses Tests Legitimate? *Journal of Business Research*, 66(9), 1261–1266.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd ed., Mahwah, NJ: Lawrence Erlbaum.

Cozby, P. C. (2005). *Methods in Behavioral Research*, 9th ed. Boston, MA: McGraw-Hill.

Crandall, C. S., & Sherman, J. W. (2016). On the Scientific Superiority of Conceptual Replications for Scientific Progress. *Journal of Experimental Social Psychology*, 66(Sept.), 93–99.

Dawson, J. F. (2014). Moderation in Management Research: What, Why, When, and How. *Journal of Business and Psychology*, 29(1), 1–19.

Demming, C. L., Jahn, S., & Boztuğ, Y. (2017): Conducting Mediation Analysis in Marketing Research. *Marketing ZFP – Journal of Research and Management*, 39(3), 76–93.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for Choosing between Multi-Item and Single-Item Scales for Construct Measurement: A Predictive Validity Perspective. *Journal of the Academy of Marketing Science*, 40(3), 434–449.

East, R. (2016). Bias in the Evaluation of Research Methods. *Marketing Theory*, 16(2), 219–231.

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.

Erfgen, C., Zenker, S., & Sattler, H. (2015). The Vampire Effect: When do Celebrity Endorsers Harm Brand Recall? *International Journal of Research in Marketing*, 32(2), 155–163.

Espinosa, J. A., & Ortinau, D. J. (2016). Debunking Legendary Beliefs about Student Samples in Marketing Research. *Journal of Business Research*, 69(8), 3149–3158.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007): G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods*, 39(2), 175–191.

Field, A. P. (2018). *Discovering Statistics Using IBM SPSS Statistics.* 5th ed. Los Angeles, CA: Sage.

Field, A. P., & Hole, G. J. (2003). *How to Design and Report Experiments*. Los Angeles, CA: Sage.

Fitzsimons, G. J. (2008). Death to Dichotomizing. *Journal of Consumer Research*, 35(1), 5–8.

Fong, L. H. N., Law, R., Tang, C. M. F., & Yap, M. H. T. (2016). Experimental Research in Hospitality and Tourism: A Critical Review. *International Journal of Contemporary Hospitality Management*, 28(2), 246–266.

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50.

Friestad, M., & Wright, P. (1994). The Persuasion Knowledge Model: How People Cope with Persuasion Attempts. *Journal of Consumer Research*, 21(1), 1–31.

Geuens, M., & De Pelsmacker, P. (2017). Planning and Conducting Experimental Advertising Research and Questionnaire Design. *Journal of Advertising*, 46(1), 83–100.

Gneezy, A. (2017). Field Experimentation in Marketing Research. *Journal of Marketing Research*, 54(1), 140–143.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing Consumer Research. *Journal of Consumer Research*, 44(1), 196–210.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than do Subject Pool Participants. *Behavior Research Methods*, 48(1), 400–407.

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis. A Regression-Based Approach*. New York: The Guilford Press.

Hayes, A. F. (2018). *Introduction to Mediation, Moderation, and Conditional Process Analysis. A Regression-Based Approach, 2nd ed*. New York: The Guilford Press.

Hayes, A. F., & Preacher, K. J. (2014). Statistical Mediation Analysis with a Multicategorical Independent Variable, *British Journal of Mathematical and Statistical Psychology*, 67(3), 451–470.

Hayes, A. F., Montoya, A. K., & Rockwood, N. J. (2017). The Analysis of Mechanisms and their Contingencies: PROCESS versus Structural Equation Modeling. *Australasian Marketing Journal*, 25(1), 76–81.

He, J., Wang, X., & Curry, D. J. (2017). Mediation Analysis: A New Test when All or Some Variables are Categorical. *International Journal of Research in Marketing*, 34(4), 780–798.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43, 115–135.

Hertwig, R., & Ortmann, A. (2008). Deception in Experiments: Revisiting the Arguments in Its Defense. *Ethics & Behavior*, 18(1), 59–92.

Hsu, D. K., Simmons, S. A., & Wieland, A. M. (2017). Designing Entrepreneurship Experiments. *Organizational Research Methods*, 20(3), 379–412.

Iacobucci, D., Popovich, D. L., Bakamitsos, G. A., Posavac, S. S., & Kardes, F. R. (2015c). Three Essential Analytical Techniques for the Behavioral Marketing Researcher: Median Splits, Mean-Centering, and Mediation Analysis. *Foundations and Trends® in Marketing*, 9(2), 83–174.

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M., & Popovich, D. L. (2015a). Toward a More Nuanced Understanding of the Statistical Properties of a Median Split. *Journal of Consumer Psychology*, 25(4), 652–665.

Iacobucci, D., Posavac, S. S., Kardes, F. R., Schneider, M., & Popovich, D. L. (2015b). The Median Split: Robust, Refined, and Revived. *Journal of Consumer Psychology*, 25(4), 690–704.

Iacobucci, D., Saldanha, N., & Deng, X. (2007). A Meditation on Mediation: Evidence That Structural Equations Models Perform Better Than Regressions. *Journal of Consumer Psychology*, 17(2), 139–153.

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental Designs for Identifying Causal Mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 5–51.

Inman, J. J., Campbell, M. C., Kirmani, A., & Price, L. L. (2018). Our Vision for the *Journal of Consumer Research*: It's All about the Consumer. *Journal of Consumer Research*, 44(5), 955–959.

Irwin, J. R., & McClelland, G. H. (2003). Negative Consequences of Dichotomizing Continuous Predictor Variables. *Journal of Marketing Research*, 40(3), 366–371.

James, W., & Sonner, B. S. (2001). Just Say Not to Traditional Student Samples. *Journal of Advertising Research*, 41(5), 63–71.

Janiszewski, C., Labroo, A. A., & Rucker, D. D. (2016). A Tutorial in Consumer Research: Knowledge Creation and Knowledge Appreciation in Deductive-Conceptual Consumer Research. *Journal of Consumer Research*, 43(2), 200–209.

Kamins, M. A, & Gupta, K. (1994). Congruence between Spokesperson and Product Type: A Matchup Hypothesis Perspective. *Psychology and Marketing*, 11(6), 569–586.

Kamins, M. A., Marks, L. J., & Skinner, D. (1991). Television Commercial Evaluation in the Context of Program Induced Mood: Congruency versus Consistency Effects. *Journal of Advertising*, 20(2), 1–14.

Kenworthy, T. P., & Sparks, J. R. (2016). A Scientific Realism Perspective on Scientific Progress in Marketing: An Analysis of Theory Testing in Marketing's Major Journals. *European Management Journal*, 34(5), 466–474.

Kline R. B. (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*, 2nd ed. Washington, DC: American Psychological Association.

Kline, R. B. (2009). *Becoming a Behavioral Science Researcher. A Guide to Producing Research that Matters*. New York: Guilford.

Koschate-Fischer, N., & Schandelmeier, S. (2014). A Guideline for Designing Experimental Studies in Marketing research and a Critical Discussion of Selected Problem Areas. *Journal of Business Economics*, 84(6), 793–826.

Kühnen, U. (2010). Manipulation Check as Manipulation: Another Look at the Ease-of-Retrieval Heuristic. *Personality and Social Psychology Bulletin*, 36(1), 47–58.

Lachowicz, M.J., Preacher, K.J., & Kelley, K. (in press). A Novel Measure of Effect Size for Mediation Analysis. *Psychological Methods*.

Leary, M. (2012). *Introduction to Behavioral Research Methods*. 6th ed. Boston, MA: Pearson.

Lehmann, S., & Bengart, P. (2016). Replications Hardly Possible: Reporting Practice in Top-Tier Marketing Journals. *Journal of Modelling in Management*, 11(2), 427–445.

Levitt, S. D., & List, J. A. (2009). Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review*, 53(1), 1–18.

Lynch, J. G., Jr. (1999). Theory and External Validity. *Journal of the Academy of Marketing Science*, 27(3), 367–376.

Lynch, J. G., Jr., Alba, J. W., Krishna, A., Morwitz, V. G., & Gürhan-Canli, Z. (2012). Knowledge Creation in Consumer Research: Multiple Routes, Multiple Criteria. *Journal of Consumer Psychology*, 22(4), 473–485.

Lynch, J. G., Jr., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015): Reflections on the Replication Corner: In Praise of Conceptual Replications. *International Journal of Research in Marketing*, 32(4), 333–342.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods*, 7(1), 19–40.

Malhotra, N. K., Nunan, D., & Birks, D. F. (2017). *Marketing Research. An Applied Approach*, 5th ed, Harlow: Pearson.

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, 9(2), 147–163.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective*, 2nd ed., New York: Taylor & Francis.

McClelland, G. H., Lynch, J. G., Irwin, J. R. Spiller, S. A., & Fitzsimons, G. J. (2015). Median Splits, Type II Errors, and False-Positive Consumer Psychology: Don't Fight the Power. *Journal of Consumer Psychology*, 25(4), 679–689.

McGuire, W. J. (1997). Creative Hypothesis Generating in Psychology: Some Useful Heuristics. *Annual Review of Psychology*, 48, 1–30.

Meyer, R. (2017). Introduction to the Journal of Marketing Research Special Section on Field Experiments. *Journal of Marketing Research*, 54(1), 138–139.

Meyvis, T., & Van Osselaer, S. M. J. (2018). Increasing the Power of Your Study by Increasing the Effect Size. *Journal of Consumer Research*, 44(5), 1157–1173.

Mill, J. S. (1843). *A System of Logic.* London: John W. Parker.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding Analysis of Covariance. *Journal of Abnormal Psychology*, 110(1), 40–48.

Morales, A. C., Amir, O., & Lee, L. (2017). Keeping It Real in Experimental Research – Understanding When, Where, and How to Enhance Realism and Measure Consumer Behavior. *Journal of Consumer Research*, 44(2), 465–476.

Ohanian, R. (1990). Construction and Validation of a Scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness. *Journal of Advertising*, 19(3), 39–52.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology*, 45(4), 867–872.

Ortinau, D. J. (2011). Writing and Publishing Important Scientific Articles: A Reviewer's Perspective. *Journal of Business Research*, 64(2), 150–156.

Perdue, B. C., & Summers, J. O. (1986). Checking the Success of Manipulations in Marketing Experiments. *Journal of Marketing Research*, 23(4), 317–326.

Peterson, R. A. (2001). On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-Analysis. *Journal of Consumer Research*, 28(3), 450–461.

Peterson, R. A., & Merunka, D. R. (2014). Convenience Samples of College Students and Research Reproducibility. *Journal of Business Research*, 67(5), 1035–1041.

Peterson, R. A., & Umesh, U. N. (2018). On the Significance of Statistically Insignificant Results in Consumer Behavior Experiments. *Journal of the Academy of Marketing Science*, 46(1), 81–91.

Pham, M. T. (2013). The Seven Sins of Consumer Psychology. *Journal of Consumer Psychology*, 23(4), 411–423.

Pieters, R. (2017). Meaningful Mediation Analysis: Plausible Causal Inference and Informative Communication. *Journal of Consumer Research*, 44(3), 692–716.

Popper, K. (1959). *The Logic of Scientific Discovery*, New York: Basic Books.

Preacher, K. J., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavioral Research Methods*, 40(3), 879–891.

Preacher, K. J., & Kelley, K. (2011). Effect Size Measures for Mediation Models: Quantitative Strategies for Communicating Indirect Effects. *Psychological Methods*, 16(2), 93–115.

Prentice, D. A., & Miller, D. T. (1992). When Small Effects are Impressive. *Psychological Bulletin*, 112(1), 160–164.

Rucker, D. D., McShane, B. B., Preacher, K. J. (2015). A Researcher's Guide to Regression, Discretization, and Median Splits of Continuous Variables. *Journal of Consumer Psychology*, 25(4), 666–678.

Sarstedt, M., Bengart, P., Shaltoni, A. M., & Lehmann, S. (2017). The Use of Sampling Methods in Advertising Research: A Gap Between Theory and Practice. *International Journal of Advertising*, https://doi.org/10.1080/02650487.2017.1348329.

Sarstedt, M., Diamantopoulos, A., & Salzberger, T. (2016a). Should we Use Single Items? Better Not. *Journal of Business Research*, 69(8), 3199–3203.

Sarstedt, M., Diamantopoulos, A., Salzberger, T., & Baumgartner, P. (2016b). Selecting Single Items to Measure Doubly Concrete Constructs: A Cautionary Tale. *Journal of Business Research*, 69(8), 3159–3167.

Sawyer, A. G. (1975). Demand Artifacts in Laboratory Experiments in Consumer Research. *Journal of Consumer Research*, 1(4), 20–30.

Sekaran, U., & Bougie, J. R. G. (2016). *Research methods for business. A skill-building approach*, 7th ed., Chichester: Wiley.

Shadish, W. R.,Cook, T. D. C., & Campbell, D. T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Shank, D. B. (2016). Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk. *The American Sociologist*, 47(1), 47–55.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, 22(11), 1359–1366.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a Causal Chain: Why Experiments are often more Effective than Mediational Analyses in Examining Psychological Processes. *Journal of Personality and Social Psychology*, 89(6), 845–851.

Spiller, S. A., Fitzsimons, G. J., Lynch, J. G. Jr., & McClelland, G. H. (2013). Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression. *Journal of Marketing Research*, 50(2), 277–288.

Sutton, R. I., & Staw, B. M. (1995). What Theory is Not. *Administrative Science Quarterly*, 40(3), 371–384.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using Multivariate Statistics*, 6th ed. Harlow: Pearson.

Till, B. D., & Busler, M. (1998). Matching products with endorsers: attractiveness versus expertise. *Journal of Consumer Marketing*, 15(6), 576–586.

Uncles, M. D., & Kwok, S. (2013). Designing Research with In-Built Differentiated Replication. *Journal of Business Research*, 66(9), 1398–1405.

Varadarajan, R. P. (1996). From the Editor: Reflections on Research and Publishing. *Journal of Marketing*, 60(4), 3–6.

Vargas, P. T., Duff, B. R. L., & Faber, R. J. (2017). A Practical Guide to Experimental Advertising Research. *Journal of Advertising*, 46(1), 101–114.

Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant Validity Testing in Marketing: An Analysis, Causes for Concern, and Proposed Remedies. *Journal of the Academy of Marketing Science*, 44(1), 119–134.

Wessling, K. S., Huber, J., & Netzer, O. (2017). MTurk Character Misrepresentation: Assessment and Solutions. *Journal of Consumer Research*, 44(1), 211–230.

Woodside, A. G. (2016). The Good Practices Manifesto: Overcoming Bad Practices Pervasive in Current Research in Business. *Journal of Business Research*, 69(2), 365–381.

Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting Researchers' Approach to Adjustment: On the Use of Covariates when testing Interactions. *Journal of Experimental Social Psychology*, 40(3), 424–431.

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research*, 37(2), 197–206.

Keywords

Experimental Process, Hypotheses, Moderation, Mediation, Manipulation.