Planning and conducting experimental advertising research and questionnaire design

Planning and conducting experimental advertising research and questionnaire design

Maggie Geuens, Ghent University

Patrick De Pelsmacker, University of Antwerp

ABSTRACT

We give an overview of the most important decisions to be taken when planning and conducting experimental advertising research. Based on previous research and state-of-the-art academic insights, we discuss good methodological practices with respect to fleshing out the contribution of a study, developing and testing experimental stimuli, selecting appropriate samples, collecting data, questionnaire design, variables and measures to be included, and scale measurement. This set of guidelines should enable experimental advertising researchers to optimize their study design and to carry it out in a correct way.

Advertising research can be divided into two streams that use a different methodology. The first approach is experimental (Eisend, 2006, 2010; Schmidt and Eisend, 2015). Well-controlled experiments are set up to test the effect of carefully designed manipulations on ad responses, and to explain why certain effects occur. The second one approaches advertising effects from a modelling angle, for instance to predict sales, or to calculate an optimal media mix (Aleksandrovs et al. 2015, Danaher and Dagger 2013, Frison et al. 2014, Naik and Raman 2003, Reynar, Phillips and Heumann 2010). In line with the latter, the increasing availability of behavioral online data has made 'big data' advertising research increasingly possible and important. The nature of big data analysis is radically different from experimental research. While experimental design focuses upon causation and explanation of phenomena, big data research focuses upon prediction, often with no or little attention to explanation and causation. Although some have announced the end of experiments, hypotheses, and theories (cf. Anderson 2008), others warn for the pitfalls of big data, such as the bias in the collection and interpretation (Bottles, Begoli and Worley 2014), the validity, transparency, replicability and interpretation of big data (Lazer et al. 2014; Boyd and Crawford 2012; Bollier and Firestone 2010), and the lack of theoretical considerations (Bottles et al. 2014). Therefore, while big data may imply a shift towards advertising research that combines big data analysis, modelling and experimentation, experimental research and small data will remain indispensable.

The purpose of this overview is to discuss the most important decisions to be taken when planning and conducting experimental advertising research. Based on previous research and state-of-the-art academic insights, we suggest correct methodological practices with respect to the different stages and components of the experimental research planning and conducting process.

As any other type of research, experimental advertising research should contribute to advertising theory and practice. This is discussed in the first section. One of the most important decisions a researcher has to take is the development of the advertising stimuli to be used in the experiments. In the second section we discuss what should be taken into account to end up with realistic and well-controlled stimuli. These stimuli are then tested in adequate samples of experiment participants. In the third section we discuss issues of sample selection such as randomness, representativeness, relevance, size, and subsample equivalence.

A large part of this overview is devoted to data collection and measurement. We propose a suitable composition and measurement sequence of an experimental questionnaire, and we extensively discuss issues such as the briefing of participants, manipulation checks, mediating and moderating variables, dichotomizing, replication factors and robustness checks, and quality control measures. Often, variables in experimental advertising research are measured by means of rating scales. In the last section, we deal with decisions to be taken when selecting or developing such scales. We discuss content and construct validity, single vs. multiple item scales, reversed item practices, response style effects, number of scale points and scale polarity, numbering and labeling. This set of guidelines should enable experimental advertising researchers to optimize their study design and to carry it out in a correct way.

To a large extent these guidelines are generic and apply to all experimental research. However, the advertising context and its constant evolution requires more attention to certain aspects of research design that are particularly relevant for advertising. For instance, stimulus and context realism become increasingly important. Advertising research should test its propositions in contexts that are ecologically valid and thus relevant for theory development and practitioners (Royne-Stafford 2016). This does not only pertain to the advertising stimuli themselves, but also to the context in which they are embedded. New advertising formats and contexts, such as in-game advertising, native advertising, content marketing, advertising integrated in social media, virtual/augmented reality ads, geo-based advertising, So-Lo-Mo promotions, etc., require more complicated and realistic stimuli that reflect the real world of advertising and its contexts. In order to make research practically meaningful, advertising researchers should not only think in terms of general theories, but also in terms of contextualization. That is, how will different target groups respond to different types of messages across different media and in different situational circumstances? (Laczniak 2015). In this respect, it is important to bear in mind that the growth of Web 2.0 has drastically changed the consumers' situational circumstances. Respondents are, for example, increasingly multi-tasking during ad exposure (cf., Angell et al. 2016, Duff and Sar 2015) and have immediate access to online reviews of advertised products (cf. Foos, Keeling and Keeling 2016).

Ecological validity also implies the selection of relevant and externally valid samples. Nowadays, lots of real-life (social media) behavioral data are available that can be used in experimental studies. However, these data are often generated in real-life situations (also

implying measures about real-life ads) hampering stimulus manipulation and sample control. Lots of experimental advertising research is carried out using online surveys with online panel participants (such as M-Turk or Prolific) which increases issues of controllability of the data collection process and 'careless responding' that should be monitored.

There is – or should be - an increasing concern about the ethicality of some (novel) advertising practices. Integrated (online) advertising formats collect data that may pose a threat to the privacy of individuals. Integrated formats such as in-game advertising, native advertising or content marketing may be considered unethical in their own right. This should be reflected in the measures that are part of advertising research that tries to explain the mechanisms behind consumers' responses to advertising stimuli, such as privacy issues, persuasion knowledge, etc. New advertising formats are increasingly aimed at behavior rather than evaluative responses such as brand attitudes. This should also be reflected in dependent variables used. Activation measures, such as eWOM intentions, liking, sharing and commenting on social media, etc. become increasingly important. These particular advertising research issues are also revisited in subsequent sections.

CONTRIBUTION TO THEORY AND PRACTICE

Although fleshing out the contribution of a study is not part of its practical operationalization, it is an important first step in the planning process. Publishing an article starts with developing a relevant research question that provides a substantial contribution, which is then investigated by means of a solid and properly conducted research design. Many papers are rejected because of lack of meaningful contribution. Unlike in many other fields, in advertising research, this contribution needs to be meaningful not only for advertising theory, but also for advertising practice. Advertising researchers thus have to think carefully before they set up a study. What will we learn that we did not know already? How does it build upon previous research and add meaningful insights to the body of knowledge about how advertising works? Why is what is studied relevant, and for whom? Why is it useful for practitioners?

These contributions need to be substantial and not incremental. One way to find that out is to conduct a thorough literature review before setting up a study. It is very sad to discover that someone else already published a similar study as the one you are conducting. Even though

the value of replication studies for the development of science is generally acknowledged and although some journals have a 'replication corner', pure replication studies do not tend to be well received by most marketing or advertising journals (Kerr, Schultz and Ling 2016). Indeed, only 2.9% of the articles published in the Journal of Advertising, Journal of Advertising Research, International Journal of Advertising, and the Journal of Current Issues and Research in Advertising from 1980 to 2012 and only 1.2% of the articles published in the Journal of Marketing, Journal of Marketing Research, and Journal of Consumer Research between 1990 and 2004 were pure replication studies (Evanschitzky et al. 2007; Park et al. 2015).

Contributions' such as 'this is the first time this is tested in Belgium', 'the authors live in Poland and Spain and therefore these two countries are studied' or 'no one has tested in-game brand placement effects in fantasy games yet' are only meaningful if the researcher can come up with a compelling argument why there is added value in the study, beyond convenience or satisfying one's curiosity. For instance, as to the fantasy game example, replicating response mechanisms that have already been studied in, say, sports games, may be theoretically and practically interesting because of the very nature of a fantasy game. These games are set in a coherent fantasy world, with myths and legends for characters (elves, trolls, …), special settings (castles, dungeons, …), specific artifacts (rings, swords, …), etcetera (Schultze and Rennecker 2007). Brands are no part of such a fantasy world. This incongruence could lead gamers to respond very differently to ads in fantasy games than to ads in sport games (Verberckmoes et al., 2016).

STIMULI

Experimental research in advertising often uses manipulated advertising stimuli to test the effect of these manipulations on consumer responses. Therefore, developing correct and relevant test stimuli is of the utmost importance.

Realism and Control

Advertising stimuli used in experiments have to strike a balance between realism and control. They have to be professionally made, of high quality, and realistic, in other words look like real advertisements. Stimulus and context realism become increasingly important due to the constant evolution of advertising and advertising contexts. Advertising research should test its

propositions in advertising contexts that are relevant for theory development and practice. New advertising formats and contexts, such as in-game advertising, native advertising, content marketing, advertising integrated in social media, virtual/augmented reality ads, geo-based advertising, So-Lo-Mo promotions, etc., require realistic stimuli that reflect the real world of advertising. For instance, testing the effects of in-game advertising realistically requires that the ad is integrated into a game (see, for instance Cauberghe and De Pelsmacker 2010). The researcher could develop static stimuli ('stills' of the game) in which in-game ads are integrated in different ways, but they would be far away from a real gaming experience, and it is then difficult to validly test the effect of, for instance, interactivity or intrusiveness. Vignettes, that is, short, carefully constructed descriptions of a person, object or situation that is shown to respondents in order to elicit judgments about these scenarios (Atzmüller and Steiner 2010), could be a valid alternative for real stimuli in case creating real stimuli is too complex. They can provide useful insights about the behavior and decisions of consumers in real-life situations (Schoenberg and Ravdal 2000). However, they are still not realistic. In sum, developing realistic stimuli may sometimes be problematic, because this can be complex and/or expensive, but realism is an important consideration in advertising research.

On the other hand, stimuli should always be developed to clearly reflect a desired manipulation. The standard procedure is to only change those elements that are needed for the manipulation, and keep the rest of the advertisement the same across manipulated conditions. Manipulation control becomes easier if this 'rest of the ad' is not too distracting or confusing. Complex, sophisticated or creative ads may well look more professional, but they can confuse the viewer and distract him or her from the components of the stimuli that are essential for the manipulation. A good stimulus for experimental advertising studies is thus well-made and realistic, but at the same time as simple as possible. Nowadays, lots of behavioral data are available that are often generated in real-life situations, with real ads (e.g., on social media). In those circumstances, control over stimulus manipulations becomes an increasingly challenging task.

In some cases, the nature of the study involves creating stimuli that contain manipulated components that are, by themselves, potentially confounding. For instance, in a cross-cultural study, Rajabi et al. (2015) manipulated the endorser of a product to be either local (an actor from Belgium, Iran or India), or global (Leonardo di Caprio). Obviously, a pre-test was used to select these endorsers, to check whether they were equally well-known in the country in

which they were used in the study. But celebrities can be different in more ways than just being famous. Prior research shows that the effectiveness of endorsers also depends on how attractive, trustworthy and credible they are (Erdogan 1999). If, say, the local celebrity is also perceived as more attractive or trustworthy than the global one, some of the effects measured may be due to these differences in attractiveness or trustworthiness, and not to the localness or globalness of the endorser. Such potential confounds have to be carefully pre-tested, and stimulus elements have to be selected that reflected the desired manipulation perfectly, without any other confounds. These potentially confounding elements should also be measured in the main experiment, so that they can be used as covariates in the analysis to neutralize their potentially confounding effect.

Real or Hypothetical Brands?

In developing realistic advertising stimuli, brand names have to be mentioned and/or brand identifiers (logos, slogans,…) have to be shown. In some cases, existing brand names have to be used for obvious reasons, for instance in brand extension studies (e.g., Dens and De Pelsmacker 2010, 2015), in studies in which one of the independents is the distinction between existing and novel brands (e.g., De Pelsmacker and Janssens 2005; Campbell and Keller 2003), in brand typicality studies (e.g., Goedertier et al. 2015), or for purposes of comparing the effects of existing vs. novel brands or of strong vs. weak brands (e.g., Dahlèn and Lange 2005). In these studies, existing brands are carefully selected on the basis of their suitability for the research objective at hand.

In many cases, however, a brand has to be used to create realistic advertising stimuli, but the brand in itself does not play a role in the study. In those cases, it is advisable to use hypothetical (new) brands, to avoid potentially confounding effects of previous exposure or experience with existing brands (Schneider and Cornwell 2005). For instance, many participants may be aware of existing brands, and may have associations with them, or positive or negative beliefs, feelings and attitudes towards them. In measuring responses to advertising stimuli, confounds may easily invalidate the results and conclusions. For instance, memory effects (e.g., brand recall) may be seriously inflated by using existing brands, as a result of which it becomes impossible to distinguish the effect of the manipulations and pre-existing brand awareness. The same goes for beliefs, feelings and attitudes. Obviously, if the same brands are used in all stimuli, that should not affect differences in effects, but effect sizes of the stimuli themselves may be flawed.

If using existing brands is deemed desirable, the best solution to neutralize the confounding effects of existing brands (or ads, for that matter) is to base the study on many (existing) ads and/or brands (see, for instance, Pham, Geuens and De Pelsmacker 2013), but this may be impractical or expensive. Another solution is to use a control group which is not exposed to any of the experimental stimuli. For instance, Dens et al. (2012) studied the effect of prominence and plot connection of eight brands placed in two movies on, amongst others, brand attitude. The existing brand attitudes for these eight brands were measured in a comparable control group that had not been exposed to the movies, and in the analysis the mean attitude score of each brand in this control group was subtracted from the post-score of the same brand in the test group to form the dependent variable. As a third alternative, sometimes pre-measures of brand attitudes in the same sample of participants are used. However, this is not an ideal practice because participants may take their pre-attitude scores into account when giving their post-scores (*interactive testing effect*, Verbeke, Farris and Thurik 1998). This could be partly avoided by also collecting pre-scores for several filler brands and/or to gather the pre-scores well before the main study is carried out, but even then the interactive testing effect may affect post-measurement. Even history effects can be induced in case an external event (repositioning, negative publicity, new campaign, …) takes place in between the pre-score measurement and the actual study.

Also new brands to be used in advertising studies have to be carefully pre-tested and selected. The variables on which hypothetical brands should be pre-tested partly depend on the context. In any case, elements such as (false) recognition and undesirable or biasing connotations should be avoided. Indeed, even new brand names or logos can evoke undesired responses or associations. For instance, a brand name can evoke a certain meaning, or people think they know the brand. This is all the more true in cross-cultural research.

SAMPLES

Randomness, representativeness and relevance

Drawing conclusions based on experimental studies is based on statistical testing. Strictly speaking, no statistical inferences can be made without the selection of a random sample out of a well-defined population, to ensure that sample characteristics differ only by chance from the population characteristics (Menon 1993). This implies that samples and subsamples should be obtained by random sampling. In advertising research, this is usually not the case,

be it because the population is not well described, or there is no sampling frame of the population, or it is impossible or cumbersome to draw a random sample. Moreover, even if it would be possible to select a random sample for a study, it would not at all guarantee representativeness of a certain population. As Shaver (1992) points out, randomness and representativeness are related only to the extent that there is repeated random sampling to ensure representativeness in the long run, not that a single random sample is wholly representative of the population.

Needless to say, the samples in most advertising studies are neither randomly drawn out of a well-specified population, nor representative of such a population. Nevertheless, in practice, they are generally accepted for theory testing, as long as they are relevant in the context of the study. Relevance of a sample means that there is a fit between the sample, the product and the consumption situation related to that product. For instance, using student or teenager samples to test the effect of public awareness campaigns against binge drinking is appropriate. For testing the effect of advertising for extending a brand of lawn mowers, a sample of people who have lived in flats all their lives is inappropriate. And using also men for testing ads for feminine hygiene products is, of course, irrelevant.

For convenience and cost reasons, many advertising studies are conducted in samples of university students. Therefore, advertising research is sometimes labeled "the science of the undergraduate marketing student" (James and Sonner 2001). Many authors have argued against using student subjects by presenting comparisons of the differences between 'college students' and 'real people'. These studies have almost universally found some differences in the results obtained from the two samples. Students are, in relevant ways, different from 'ordinary' consumers. They usually have less money, their life style is atypical for the consumption situation of most consumers, and they usually have no or less experience in buying most consumer products. This may easily bias their perceptions, motivations and preferences. Some authors have argued that these differences are small enough to justify using students as research subjects, but most, however, conclude that students are not good surrogates for adult consumers (for a discussion and overview, see James and Sonner 2001). Calder et al. (1981) make the distinction between what they call "effects application" and "theory application". The former aims at statistical generalization of a theory. In that case, a close correspondence between the research sample and the population, and statistical sampling, is required. Student samples would often be inappropriate in this case. The latter

aims at theory falsification, for which any (preferably homogeneous) sample in the theory domain is suitable. In this case, student samples could be used.

Besides student samples, researchers increasingly make use of samples drawn from Crowdsourcing Websites, of which MTurk is the largest and the one most often used (Shank 2016). MTurk offers clear advantages such as a low cost, easy handling, a fast turnaround time, a large and heterogeneous respondent pool, respondents from different nationalities (allowing cross-cultural comparisons), respondents from hard-to-reach target groups, a good response rate for follow-up studies, truly random allocation to experimental conditions, and importantly, reliable, consistent, and high quality data (Burmester, Kwang and Gosling 2011; Berinsky, Huber and Lenz 2012; Mason and Suri 2012; Rand 2012; Shank 2016). However, although the MTurk population is diverse, it cannot be considered as representative of the entire population. In general, MTurk respondents are younger, more educated, less employed, less religious, more liberal, have a lower income, and consist of relatively more females than the general US population (Berinsky et al. 2012; Mason and Suri 2012; Paolacci and Chandler 2014, Paolacci, Chandler and Ipeirotis 2010; Shank 2016).While results obtained with an MTurk sample appears to correspond well with other online and offline samples (Shank 2012), "Super-Turkers" (i.e., quasi-professional respondents with excessive experience) and "Spammers" (i.e., respondents only interested in maximizing their pay rate) represent a major concern (Deetlefs, Chylinski and Ortmann 2015). Deetlefs et al. (2015) recommend to take in a self-report question on experience and several control questions to enable researchers to discard both types of respondents (we discuss this further in the section 'Quality Control'). If these issues are taken care of, MTurkers usually make a good sample for experimental research.

In any case, researchers should keep in mind that ensuring some form of representativeness and randomness is desirable, and relevance is imperative.

Sample Size

How large should subsamples per experimental condition be? A study must be of adequate size. It must be 'big enough' to find effects of scientific relevance to be also statistically significant. However, it should also be not 'too big' such that an effect that is not scientifically relevant becomes statistically significant (Lenth 2001). Many experiments in consumer behavior and advertising are conducted in small groups, often even as small as 20

participants or less per experimental condition. As argued in the next section, small subsamples can be problematic in terms of equivalence, but they also lead to less power in statistical testing. Formally, the power or sensitivity of a test is the probability that the test correctly rejects the null hypothesis (H_0) when the alternative hypothesis (H_1) is true. In other words, it is the probability of accepting the alternative hypothesis (H_1) when it is true, that is, the ability of a test to detect an effect, if the effect actually exists. Weak power leads to type II errors: existing scientifically meaningful effects cannot be confirmed. Consequently, differences between scores in different experimental conditions will have to be quite substantial before they are statistically significant. On the other hand, large subsamples lead to higher statistical power, and even relatively small and scientifically insignificant differences between scores may be statistically significant. However, to be meaningful, differences should not only be statistically significant, they should also be substantial enough to warrant conclusions about the acceptance or rejection of hypotheses. In other words, too much power can lead to Type I errors: a non-existing effect is found to be statistically significant.

Typically, in experimental advertising research, one wants to test the significance of the difference between scale scores of different experimental conditions. Often, this difference is relatively small, so sufficiently large subsamples are needed to achieve sufficient statistical power. For instance, for a test comparing two independent samples (as in a simple effects test in a between-subjects design) in which the scores on a construct are 3.0 and 3.6 (SD=.80) on a 7-point scale, the significance level required is .05, and the required statistical power is 80%, the required size for each subsample is 28. The higher the desired power, the smaller the difference in scores, and the larger the standard deviations, the larger the samples need to be. For instance, in the given example, if the SD = 1.00, a sample size of 41 is needed (www.stat.ubc.ca/~rollin/stats/ssize/n2.html). Obviously, the researcher does not know what the scores and their standard deviations will be. However, to ensure sufficient statistical power, a size of 30-40 participants per experimental condition seems to be indicated.

If in the study a number of variables are measured that will be used as moderators in the analysis, larger subsamples per experimental condition are needed. For instance, assume that the researcher sets up a two-condition experiment. In one condition, participants are exposed to an advertisement for a new product in which a celebrity is used as an endorser, and in the second condition the participants see an advertisement for the same product with no visual, but with a lot of product information. A total sample of 60 would be sufficient to test the

differences in responses between the two conditions (30 per condition). However, if during the experiment, participants are asked for their gender, and in the analysis the researcher wants to test whether gender moderates the effect of type of stimulus on the attitude towards the product (a gender (2) x stimulus type (2) interaction), then 60 participants will not be enough, because in terms of analyses, there are now four conditions instead of two, and simple effects testing within the interaction would be based on subsamples of, on average, 15 participants.

Subsample Equivalence

What is extremely important for a valid experimental design, is the comparability or equivalence of samples across experimental conditions. The traditional solution for this is to randomly assign participants to experimental conditions, but that may not be enough. Especially in small (sub)samples, randomness does not necessarily lead to sample equivalence, especially in populations that are heterogeneous in terms of important and relevant characteristics in the context of the study. For instance, if the study is done in a sample of business undergraduates (assuming that is a relevant sample for the topic under study), the participants will at least be relatively homogeneous in terms of age, level of education, knowledge about business administration, and certain life style elements. In this case, random allocation of participants to experimental conditions may lead to sufficient subsample equivalence. However, if the population of interest is very heterogeneous in terms of age, level of education, profession, etc., and all these are relevant characteristics in the context of the study, than randomly assigning individuals to small experimental subsamples will not necessarily lead to subsample equivalence in terms of several of these characteristics.

There are three solutions for this problem. The first one is to increase sample sizes. The second one is to select a stratified sample. That is, define categories of participants on the basis of variables that are relevant in the context of the study, decide which part of the samples should be in which category, and then allocate participants accordingly. In face-to-face or computer-aided lab experiments, this can be easily organized, but also in online survey tools, this can be easily programmed. The third is to make use of matched samples. That is, if a higher educated, middle-aged woman is assigned to one condition, a similar higher educated, middle-aged woman should be assigned to the other condition(s).

Nowadays, lots of data that is generated in real-life situations are available that can be used in experimental studies. In such circumstances, control over the composition and equivalence of samples across experimental conditions becomes increasingly challenging.

DATA COLLECTION

After carefully crafting stimuli and/or primes and selecting test samples, participants have to be exposed to the stimuli, and their responses and characteristics have to be measured. Research results can be seriously confounded by the use of inappropriate measurement procedures. Therefore, the content and sequence of the measurements is very important. Table 1 outlines a recommended sequence of measures in a questionnaire.

PLACE TABLE 1 ABOUT HERE

Briefing

Upon arrival in the lab or when starting to fill out a questionnaire, respondents should receive a proper introduction or briefing. This briefing usually entails some form of deception to conceal the real goal of the research (a cover up story). Misleading respondents is often necessary in order to avoid demand effects. That is, if respondents are aware of the researchers' objectives, some may decide to counteract and cross these objectives while others may take on the role of a 'good subject' and behave in line with the researchers' hypotheses (e.g., Orne 1962; Rosnow and Rosenthal 1997). Either way, no true response to the manipulated stimuli will be obtained. In order to enhance the credibility of the cover up story, researchers often take in filler items and/or rely on the '*independent studies paradigm'*. The latter refers to having respondents believe that they are participating in several, unrelated studies. Tessitore and Geuens (2013), for example, disguised that they investigated the effectiveness of product placement by telling respondents that they would participate in two separate experiments. The so-called first experiment was allegedly from researcher x who was interested in reactions of viewers to specific movie fragments, whereas the so-called second experiment was from researcher y who was interested in the strength of different brands.

Notwithstanding using a good, believable cover up story, at the end of the study a *suspicion probe* should check whether respondents believed the ostensible purpose of the study. The use of such a probe enables researchers to identify (and possibly eliminate) those participants who

figured out the real purpose (Taylor and Shepperd 1996). After the suspicion probe, some form of debriefing is necessary, especially when participants have been given false information as part of the briefing, or when participants' psychological states have been manipulated. In a debriefing, the researchers explain the true purpose of the study to the participants and explain when and how they were manipulated.

While demand effects may never be completely ruled out, they tend to pose a bigger problem for within-subjects designs than for between-subjects design (Pany and Reckers 1987). In a between-subject design, participants are only exposed to one manipulated stimulus. In a within-subjects design, participants are exposed to several stimuli. Researchers sometimes decide to use a within-subject design because it requires less respondents and has more power as individual differences are controlled for (Kirk 1982). However, exposing respondents to multiple treatment levels of an experimental factor (e.g., in an effort to investigate the effectiveness of a celebrity endorser, show respondents an ad for brand x with an expert endorser as well show them an almost identical ad for the same brand with a non-celebrity endorser) makes it very easy for respondents to figure out what the real purpose of the study is.

Manipulation (Check)

The briefing introduction is usually followed by exposing the participants to a manipulation. This manipulation is often embedded in the advertising stimulus itself, but it can also consist of a prime or priming instruction. For example, to investigate the role of construal level on ad persuasiveness, researchers could manipulate the ad in such a way that it evokes an abstract vs. concrete construal (e.g., by referring to the respondents' home country vs to a distant country) or they can prime spatial distance externally by having respondents describe their opinion about getting a degree in their home country vs. a distant country before they are exposed to the focal advertising stimulus (see for example, Kulkarni and Hong 2015).

After exposure to the manipulated stimulus, the dependent variables are measured, such as feelings, thoughts, attitudes towards the stimulus and the brand, and behavioral intentions. It is advised to order these dependent variables in reverse-causal order, especially if the researchers are interested in the relationship between these variables, for example, by first measuring behavioral intentions, than brand attitudes, and then stimulus-related responses such as the attitude towards the ad, feelings and thoughts.

Irrespective what type of manipulation is used, to avoid any bias a manipulation check can cause, it is better to insert the manipulation check after measuring the dependent, mediating, moderating and control variables. Some researchers choose to have a manipulation check immediately after the manipulation, but this is usually not a good idea. For example, if a researcher is interested in the effectiveness of humor in advertising, probing to what extent respondents thought the ad was humorous before measuring the dependent variables (often attitudes and intentions), could easily induce demand effects, but could also make respondents take into account the humorous aspect to a larger extent than they would if their attention was not drawn to it. Prior research shows that context factors, such as manipulation checks, alter the relation between variables (Richter 2010). For example, when investigating ease-of-retrieval effects (that is, the effect experienced ease by which information can be brought to mind has on attitudes, judgments, and preferences, cf. Schwarz et al. 1991a), studies that found evidence for ease-of-retrieval effects under low elaboration conditions always included the manipulation check before the dependent variables. The studies that measured the manipulation check after the dependent variables did not find evidence of ease-of-retrieval effects in this same elaboration condition, the reason being that, if the manipulation check preceded the dependent variables, this procedure increased the likelihood that respondents relied on ease-of-retrieval when making their judgment (Kühnen 2010). For the same reason, it is also advised to measure mediating, moderating and controlling variables after the dependent variables (Kühnen 2010).

In case the time lapse between exposure to the manipulation and the manipulation check is quite large, the stimulus or prime can be shown again to avoid that respondents do not clearly remember it anymore. Another possibility is to not use a manipulation check in the main study, but extensively pre-test the manipulated stimuli or primes in a separate study, for instance when manipulation checks in the main study can be confounded by previous measures or questions. However, in the latter case, the researcher never knows for sure that the manipulation also worked properly in the main study. Also, using a manipulation check in the main study, provides the researchers the possibility to discard those respondents from the analyses for whom the manipulation was clearly ineffective.

Which elements of a manipulation should be tested? To be on the safe side, a conservative rule would be to use manipulation checks (or pre-test manipulations) whenever there is doubt

about how ‘obvious’ manipulations are. Manipulation checks are not needed for obvious and blatant facts, such as colours, man/woman, numbers (e.g. number of arguments or number of people in an ad).

Mediation: Measurement-of-Mediation and Experimental-Causal-Chains

The basic premise of experimental advertising research is that it tests the relationship between one or more predictor (independent) variables (e.g., different advertising stimuli) on one or more criterion (dependent) variables (e.g., brand attitude, purchase intention). However, behind this straightforward logic, there may be more than meets the eye.

Behind the relation between a predictor and an outcome (criterion), there may be a mechanism that “explains” the effect. For instance, personalized advertisements (predictor) lead to a better attitude towards the ad, because personalized advertisements are perceived as more relevant (De Keyzer, Dens and De Pelsmacker 2015). In other words, relevance “explains” the relation between personalization and brand attitude. In this example, “relevance” is a mediator. A mediator accounts for the relation between the predictor and the criterion. Mediators explain how a predictor influences a criterion. They clarify what would otherwise remain a black box in terms of why a manipulated stimulus predicts an outcome. Consequently, experimental advertising research should include potential mediators. They are important for theory building.

If mediators are measured variables, Hayes’ PROCESS macros have become the standard approach to test the (serial) mediation process (Hayes 2013). However, Spencer, Zanna and Fong (2005) warn against the use of measurement-of-mediation designs. They should only be used when the proposed underlying process (i.e., the mediator) is easy to measure and difficult to manipulate. A more powerful methodological approach to assess mediation is the use of experimental-causal-chain designs, that is a series of experiments that demonstrate the proposed causal chain. For example, to provide evidence that downward versus upward head movements increase preference-decision consistency because looking down (vs. looking up) induces a concrete construal level (vs. abstract construal level), Van Kerckhove, Geuens and Vermeir (2015) showed that (1) looking down vs. looking up leads consumers to adopt a lower (higher) level of construal and (2) priming a lower (higher) level of construal increases (decreases) preference-decision consistency. Spencer, Zanna and Fong (2005) argue that

experimental-causal-chain designs are recommended whenever the mediator is easy to manipulate. The reason is that manipulating both the independent variable and the mediating variable allows to make strong inferences about the causal chain of events.

Moderation and Median Splits

The relationship between a predictor and a criterion is also subject to boundary conditions. What may be a strong and positive relationship in some circumstances, may be a weak or negative relationship in other circumstances. For instance, highly anxious people are strongly influenced by fear appeals, whereas less anxious people are less affected (De Meulenaer, Dens and De Pelsmacker 2015). In this example, anxiety is a moderator. A moderator affects the direction and/or strength of the relation between a predictor variable and a criterion variable. In experimental research, moderator effects can be represented as an interaction between an independent variable and a factor that specifies the conditions for its operation. Moderators specify when and to what extent certain relations between an independent (predictor) and a dependent (criterion) will hold. Consequently, also moderators are crucial for theory development and for practice. They nuance a relationship and they shed light on the boundary conditions of an effect.

A moderator can be a qualitative (e.g., sex, race, class) or quantitative (e.g., level of anxiety) variable. Qualitative moderators can be easily added as an extra factor in ANOVA analyses. However, also for quantitative moderators it used to be standard procedure to median split the scores on these moderators, and use the dichotomized variable as an extra factor in ANOVA analyses. Although the question of median splitting a continuously measured variable is an analytical decision and, as such, outside the scope of the present overview, it has to be taken into account in setting up an experiment.

The question as to whether it is good practice to median split or dichotomize a continuous variable for further analysis, has been a matter of debate for a long time. It is common knowledge that dichotomizing a continuous variable leads to loss of information (individual scores are dichotomized) and loss of statistical power. Consequently, median splits lead to greater type II errors: scientifically meaningful effects will not be statistically significant. Some scholars argue that this is not a big deal. They argue that it is not problematic for an effect not to be found, it just makes studies more conservative. Moreover, median splitting does not produce spurious results (type I errors). That is, effects that are not there will not be

statistically significant (Maxwell and Delaney 1993). Median splits, they further argue, are warranted if the researcher is mainly interested in group effects rather than individual effects. Finally, analyses performed on categorized data, typically ANOVAs, would be easier to interpret and understand than moderated regression analysis typically used on the combination of manipulated factors and continuous moderators (Iacobucci et al. 2015).

However, the bulk of methodologists have argued against median splits for quite some time (e.g. Fitzsimons 2008; Irwin and McClelland 2003). First of all, they argue that dichotomization discards the potentially rich variation in individual scores. For example, subjects scoring just above the median, moderately above the median, and substantially above the median are all treated as identical. Importantly, they argue that, in many cases, median splitting does increase the risk of type I errors (Rucker, McShane and Preacher 2015). Others also argue that loss of statistical power due to median splitting a continuous variable is not a decision to be "conservative", but rather a decision to be less persuasive: more statistical power leads to bigger effect sizes and more persuasive results (McClelland et al. 2015). All in all, there is a growing consensus that moderators should not be dichotomized, and using continuous measures to test moderation is the best research practice.

Moderated regression analyses, using a combination of categorical manipulated factors and measured predictors and moderators has become the normative standard of analysis. Hayes' PROCESS macros have become the standard approach to moderation analysis. To interpret the results, "spotlight" analyses (Aiken, West and Reno 1991) or "floodlight" analyses (Spiller et al. 2013) can be used.

Moderating and mediating variables often constitute the real contribution of a study.

Replication and Robustness Check

From conventional statistical knowledge we can infer that, with a conventional significance level of 5%, one out of twenty times a true null hypothesis will be rejected (Type I error). Taking into account that for most studies, multiple statistical tests are carried out, that authors usually do not report all the analyses run, and that papers with significant findings are more likely to be submitted to and accepted by journals, the likelihood on Type I errors in published papers is even much higher than the 5% statistical tests would suggest (Eisend, Franke and Leigh 2016). Therefore, to be able to trust the findings of a study, replication of the results in

follow-up experiments is crucial and should be an integral part of each research design (Uncles and Kwok 2013).

Next to replicating the main findings in a different sample and/or using different stimuli, follow-up experiments can be designed in such a way that they rule out confounds or rival theoretical explanations, and allow to investigate mediating and/or moderating processes. For example, to provide robust proof that a symbolic vs. physical product presentation leads consumers to purchase less unhealthy products in an online than in a brick-and-mortar store, Huyghe et al. (2016) set up a series of experiments in which each experiment replicated the main effect (i.e., consumers buy less vices online than offline), but ruled out other explanations and confounding variables such as (1) being exposed to the unhealthy products or not, (2) store atmospherics, (3) payment method, (4) order lead time, and (5) shopping motive.

As mentioned, next to mediating and moderating variables, potential confounding effects should be controlled for. As such, variables like age, gender, mood, brand/product familiarity, product/familiarity usage, etc. are often measured and included in the analyses as covariates. Control variables should only be included in the analyses if there are solid theoretical or logical reasons, or if previous research provides indications to do so.

Quality Control

Data collection in experimental studies are done in face-to-face settings (be it paper-and-pencil or in a computer lab) or, increasingly, online. Online studies usually have the advantage of being fast, cheap, and involving a more representative sample. Furthermore, as a self-administration technique, it offers respondents the possibility to work at their own pace and it avoids the presence of an interviewer (and a corresponding interviewer bias) (e.g., Catania et al. 1996, Davis et al. 2010). Importantly, the presence of an interviewer may lead respondents to answer questions in a different way, for example, by expressing a clearer or more extreme opinion (Jordan, Marcus and Reeder 1980; Weijters, Schillewaert and Geuens 2008). Interviewer presence may also generate less honest answers as it increases respondents' inclination to impression management (i.e., social desirability) (Krumpal 2013; Paulhus 2003). To control for social desirable responding, it may be wise to always include a scale such as the Balanced Inventory for Desirable Responding (see Steenkamp, De Jong and Baumgartner 2010 for more details).

On the other hand, researchers should realize that they have less control over the conditions in which respondents fill out an online questionnaire, due to the physical distance, lack of personalization and distraction. In a face-to-face setting, participants can be briefed and monitored in real time. Lab experiments may therefore be better to preserve internal validity. For example, to investigate the impact of consumers' mood on their attitude-behavior consistency, Elen et al. (2001) manipulate mood by exposing respondents to either a negative or a positive movie clip before they have to make choice decisions. If such a study is run online, respondents that take a break in between the mood induction and completing the dependent variables become useless as the mood manipulation will not last long enough to affect the dependent variables. Also, multi-tasking (e.g., listening to music or watching TV while completing the questionnaire), the presence of friends or relatives, distracting or annoying background noise, etc., may all render manipulation efforts completely ineffective. In a recent survey MTurkers, for example, admit to frequently multitask with 18% of them watching television, 14% listening to music and 6% instant messaging while filling out a questionnaire (Chandler, Mueller and Paolacci 2014).

Irrespective of whether the study is run online or in a lab, a number of controls should be built in to identify and eventually remove careless or inattentive respondents. This is important, because data from such responses can seriously invalidate the results of a study. Various control mechanisms have been proposed (see, for an extensive overview, Mead and Craig, 2012). For instance, most questionnaire software allows to set minimum and maximum duration times for exposure to stimuli, to ensure that participants do not click away the stimulus too quickly. Most questionnaire software also allows to measure the time that participants spend on each page and how long it takes them to complete the full questionnaire. The researcher can then decide later on to discard participants that completed the survey too quickly or took a break after the experimental manipulation. For example, Deetlefs et al. (2015) excluded respondents that filled out questions faster than readable, as well as respondents that needed time more than three standard deviations away from the mean.

Another strategy is to try to prevent inattentive responses via instruction sets. However, researchers should realize that not all participants are diligent in reading and following instructions. Participants who do not read the questions and/or fail to follow instructions induce noise and threaten the validity of the data. In order to be able to identify such participants, Oppenheimer, Meyvis and Davidenko (2009) suggest to insert an instruction

manipulation check. That is, a question embedded within the other questions, which is similar in length and response format (e.g., also a 7-point Likert scale). However, in contrast to the other questions for which respondents have to report their own opinion, they are asked to follow an instruction, in order to identify respondent care in response, or to flag those that are not carefully reading the item. Examples include social desirability and nonsensical or "bogus" instructed response items (e.g., "if you are completing a questionnaire now, respond with a six for this item"; "To monitor quality, please respond with a two for this item"). Another option is to include self-report measures of response quality placed at the end of a survey ("In your honest opinion, should we use your data?") (Mead and Craig 2012). In addition, control questions can be included that probe what the stimulus, scenario or instruction was about to get an idea of which respondents failed to pay sufficient attention. Respondents who fail these checks better be excluded from further analyses.

There are also several indices that can be computed post hoc for identifying careless response (Mead and Craig 2012). *Consistency indices* match items with control items that are highly similar, be it in meaning, or based on known empirical correlations among these items. A lack of consistent responding is then indicated by deviation among responses to similar items. *Response pattern indices* identify persons responding too consistently to items measuring theoretically distinct constructs. These indices are typically computed by examining the number of consecutive items for which a respondent has indicated the same response option. Excessive utilization of a single response option can flag careless responding. Again, these respondents should be removed from the analysis. Finally, if a respondent pool is used of which some respondents may have excessive experience with filling out questionnaires (cf. super-Turkers), it is advisable to include a question that probes how often respondents participate in academic research. Also these over-experienced respondents would better be discarded from the analyses (Deetlefs et al. 2015).

MEASURES

When designing the concrete measures for their variables, researchers have to make several decisions, both concerning the items that will be taken in and the scale format that will be used to measure these items. We discuss the most important issues related to each of them below.

Decisions Concerning the Items

Content or Construct Validity

Researchers should pay careful attention to only use measures that adequately represent the construct, and thus possess content or construct validity (Nunnally 1978; Rossiter 2002). Put differently, it is important to use measures that really measure what they intend to measure. Loose definitions and operationalizations of constructs, which embrace several other characteristics besides the intended ones or that do not embrace all aspects of the construct, induce a construct validity problem and leave the researchers and the readers uncertain of what was actually measured. For example, recognizing that consumers may have different motivations to purchase innovative products but that most scales take into account two different types of motivations at most, Vandecasteele and Geuens (2010) developed a new consumer innovativeness scale that fully grasps the four types of motivation (functional, hedonic, social, and cognitive) that underlie consumer innovativeness.

According to Rossiter (2002), in order to achieve content or construct validity, a construct should be defined in terms of an object, attribute and rater entity. Take as an example how Zaichkowsky (1994) defined 'involvement' for her 'personal involvement inventory', that is "[a] person's perceived relevance of the advertisement" (p. 61). The object here is 'the advertisement', the attribute is 'personal relevance', and the rater refers to 'a person', or more specifically, a consumer. Personal relevance, though, had better be further defined in terms of its underlying dimensions as Zaichkowksy conceptualized it as consisting of a cognitive and affective dimension (Rossiter 2002).

Fortunately, over the past decades many marketing and psychological scales have been extensively validated and large scale inventories have been built from which researchers can easily select scales (e.g., Bruner 2015; Bruner, James and Hensel, 2001; Bearden and Netemeyer 1999). In case no validated scales are available, researchers can turn to the well-established methodological guidelines to construct their own scale (Churchill 1979; Peter 1981; Rossiter 2002). But, as mentioned, even when they turn to existing scales, researchers should ascertain that the measure is construct valid.

Single vs Multiple Items

In his influential article, Churchill (1979) states that marketers should develop and use multi-item scales to measure constructs. Since then, standard practice in (experimental) advertising research is to measure often used constructs such as the attitude towards the ad (Aad) or the brand (Ab) by means of multi-item scales. To be more precise, researchers have used multiple items to measure the attribute of the construct (e.g., attitude, quality, liking), for the object of the construct (e.g., a company, a brand, an ad). Multiple-item measures are assumed to be inherently more reliable than single-item ones because they enable computation of correlations between items. If these correlations are positive and produce a high average correlation (i.e., a high coefficient alpha), this indicates the 'internal consistency' of all the items in representing the presumed underlying attribute (Bergvist and Rossiter 2007). Moreover, a multiple-item measure captures more information than can be provided by a single-item measure and a multiple-item measure "is more likely to tap all facets of the construct of interest" (Baumgartner and Homburg 1996, p. 143). However, practitioners favor single-item measures, on the practical grounds of minimizing respondent refusal and cost.

On the basis of comparative predictive validity tests, Bergkvist and Rossiter (2007) argue that Aad and Ab could be measured by means of single-item measures, and that this could also be the case for all measures that are "double concrete". Double concrete means that, first of all, the object of the construct (e.g. an advertisement, a brand, a company) is 'concrete singular', meaning that it consists of one object that is easily and uniformly imagined. Second, the attribute of the construct is 'concrete', again meaning that it is easily and uniformly imagined, such as Aad and Ab. Another argument for single-item measures is the desire to avoid common methods bias. Common methods bias occurs when the correlation between two or more constructs is inflated because they were measured in the same way (see, e.g., Podsakoff et al. 2003; Williams, Cote and Buckley 1989).

However, to measure abstract constructs validly, multiple items are needed. A construct is "abstract" if (1) the object of the construct consists of two or more components (e.g., the materialism value, which has three components—namely, use of possessions to judge success, centrality of possessions in a person's life, and the belief that possessions lead to happiness), or consists of a set of constituent sub-objects (e.g., for job satisfaction, aspects of a person's job, such as supervisor, coworkers, job duties, workplace technology, and policies or (2) the attribute of the construct is formed from two or more components (e.g., service quality, with

its components of reliability, responsiveness, empathy, and so forth) or elicits and is reflected in a series of mental or physical activities (e.g., most personality states or traits).

Despite 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.

Reversed Items

Should researchers include reversed items (i.e., items that need to be recoded to show a relation in the same direction with the underlying construct) in their scales or not? This question is not easy to answer as the practice has clear advantages, but also some disadvantages. The use of reversed items can enhance respondents' attentiveness (Barnette 2000), it can increase construct validity as the construct may be more fully grasped (Carroll et al. 1999), it can counter respondents' tendency to agree (i.e., acquiescence bias) (Paulhus 1991), and it can lower respondents' confirmation bias (in the sense that the inclusion of reversed items makes it harder for the respondents to guess what the researchers are exactly after (Weijters and Baumgartner 2012)). However, the use of reversed items can also confuse respondents and often goes together with methodological problems such as low factor loadings and poor internal consistency because of their weaker correlation with the non-reversed scale items (e.g., Motl and DiStefano 2002; Quilty, Oakman and Risko, 2006).

In as far as reversed items are obtained by using negations (e.g., by inserting the particle 'not' as in 'Most advertising is believable' – 'Most advertising is not believable'), research shows that such negations are especially confusing (Swain, Weathers and Niedrich 2008) and thus should better be avoided. In contrast, reversed items can also be obtained by using an antonymic expression (e.g., 'advertising usually is believable' – 'advertising often is misleading'). According to Weijters and Baumgartner (2012), a careful use of such reversals can be beneficial because simply leaving out reversed items still leaves the researcher with the problem of careless responding, acquiescence, and confirmation bias. When no reversed items are used, the method effects resulting from these biases cannot be disentangled from content and as such become harder to detect.

Then, how should researchers proceed with reversed items? First, extremely worded reversed items (such as 'I love this ad' – 'I hate this ad') should be avoided since they lead to

inconsistent responses. Second, it is advised to use balanced scales (that is, an equal number of normal and reversed items) to avoid response inconsistencies that may arise when respondents with a positive answer on an item do not have an opposite response option to choose from for the reversed item. Third, to keep respondents attentive and to avoid wrong expectations, it may be best to alternate the keying of the items. Finally, since research shows that respondents consider a wider variety of relevant beliefs the further apart related items, it is best to distribute related items throughout the questionnaire, separated by unrelated buffer items (Weijters and Baumgartner 2012; Weijters, Geuens and Baumgartner 2013, Weijters, Schillewaert and Geuens, 2009).

Decisions Concerning the Scale Format

Constructs such as attitude towards the ad, attitude towards the brand, purchase intention, involvement, and several behavioral responses are usually measured by means of rating scales. These rating scales come in different formats and differ, for example, in terms of the number of scale points, the valence of the numbering, scale polarity and the way scale points are labelled. The choice of a scale format influences the way respondents answer the questions (Tourangeau, Rips and Rasinski 2000). The scale format can bias means, making ad responses look more or less positive than they really are (Weijters, Cabooter and Schillewaert 2010), it can affect item variances affecting how homogeneous or heterogeneous participants' ad responses seem (Greenleaf 1992), and it can alter correlations between items and constructs (Baumgartner and Steenkamp 2001). However, even though the choice of the scale format alters research results and conclusions, researchers often do not contemplate on or motivate why a certain format is chosen for. They even regularly use a scale format that deviates from the originally validated scale format (Cabooter et al. 2016).

Below, we summarize some of the most important research results and repeat some of the guidelines offered in the literature. However, we first pay attention to response styles, as they are the main source of the contamination of research results. Response styles can emerge no matter which scale format is used (as they are also respondent- and context-dependent), but some scale formats are more vulnerable than others.

Response Style Effects

Respondents' answers on questionnaire items are not always a reflection of their true opinion due to the presence of measurement errors such as response styles (Paulhus 1991; Baumgartner and Steenkamp 2006). Response styles can be defined as a tendency to select a certain response category of a rating scale more often regardless of the item content (Paulhus 1991). Frequently studied response styles are acquiescence (a preference for the positive response options), disacquiescence (a preference for the negative response options), extreme responding (a preference for the extreme response options), and midpoint responding (a preference for the middle response option) (Baumgartner and Steenkamp 2001).

Response styles affect the validity of research conclusions as they affect item means and variances, as well as the correlations between items. For example, the presence of acquiescence (disacquiescence) inflates (deflates) item means and inflates (inflates) the estimated relation between items (Baumgartner and Steenkamp 2001). Response styles can be measured and corrected for, but this requires that researchers include a set of heterogeneous items that share no common content neither with one another nor with the substantive items (Weijters, Baumgartner and Geuens 2016). Such a set of unrelated, heterogeneous items is needed in order to be able to disentangle response styles from content. For example, if researchers only include substantive items such as scales measuring respondents' attitude towards a specific ad, attitude towards the promoted brand, and purchase intention, responses on these scales likely are highly correlated. Therefore, if a respondent rates each item very positively, it is impossible to know to what extent these responses reflect the respondents' true opinion and to what extent they are a result of a response style. Therefore, questions unrelated to the ad and the specific brand should be included as well, such as, for example, questions on sports, politics, gender equality, charity, etc. For a set of k unrelated, heterogeneous control items measured on a 7-point response scale (1=disagree, 7=agree), acquiescence, disacquiescence, extreme responding and midpoint responding can be measured as follows (Baumgartner and Steenkamp 2001):

$$\text{acquiescence} = [f(7)*3 + f(6)*2 + f(4)]/k$$
$$\text{disacquiescence} = [f(1)*3 + f(2)*2 + f(3)]/k$$
$$\text{extreme responding} = [f(1) + f(7)]/k$$
$$\text{midpoint responding} = f(4)/k$$

where f (x) refers to the frequency a respondent selects response option x.

Several options are available to correct for response styles (Weijters, Baumgartner and Geuens 2016). For example, the original data can be corrected for response styles by regressing the substantive items on the computed response style measures and using the residuals from this regression in subsequent analyses (Baumgartner and Steenkamp 2001). An alternative to computing and working with residuals is to use the response style measures as covariates in the main analyses. In this respect, the Representative Indicators Response Style Means And Covariance Structure (RIRSMACS) approach, in which each substantive item is specified as a function of multiple measures of each response style, proves to be a reliable, but also more complicated method (Weijters, Schillewaert and Geuens 2008).

The foregoing methods correct for response styles at the individual level. However, sometimes it may be necessary or useful to correct on a group level. This could be the case in cross-cultural research when different languages or different nationalities lead to differences in how scales are used (e.g., Weijters , Geuens and Baumgartner 2013; see also Puntoni, Weijters and Baumgartner in this issue), but also in experimental research if it can be assumed that the experimental conditions trigger differences in scale usage. For example, as extreme responding has been shown to be more prevalent the more respondents are involved with the topic (Gibbons, Zellner and Rudek 1999), an involvement manipulation may induce response styles at the group level (i.e., on the level of the different conditions). The same goes for a cognitive load manipulation as cognitive load induces more acquiescence (Knowles and Condon 1999). An easy and convenient way to correct for group differences in response styles is the recently developed calibrated sigma method (Weijters et al. 2016). Whereas for a 5-point Likert scale, researchers usually assign "1" to the 'completely disagree' category, "2" to the 'disagree' category, "3" to the 'neutral' category, "4" to the "agree" category and "5" to the 'completely agree' category, the calibrated sigma method assigns different numbers to these categories for different groups (conditions). More specifically, the chosen numbers depend on how often the different response options are chosen for a set of heterogeneous, control items by the participants in a specific group. Thus, instead of automatically assigning the scores 1 to 5, the scale scores depend on the response behavior of the different groups. As such, 'agree' is not coded as 4 for the whole sample, but might be coded as 4.2 for one group and 3.8 for another group (Weijters et al. 2016).

It has to be mentioned though that, despite the general recognition that response styles threaten the validity of research findings, they are seldom corrected for. This is probably due

to the fact that a set of extra items needs to be included in the questionnaire or that some correction methods are rather sophisticated. The more controlling for response styles is difficult, the more important it becomes to keep the occurrence of response styles as small as possible.

Number of Scale Points

In deciding on the number of scale points, researchers have to reconcile two considerations: the more scale points, (1) the finer the information that can be gathered, but also (2) the more complex and time consuming the choice task becomes as respondents must discriminate between more and finer response categories (Weathers, Sharma and Niedrich 2005).

A general recommendation in the literature is to use between five and nine scale points (Cox 1980; Schwarz et al. 1991b). Weijters, Cabooter and Schillewaert (2010) observed that seven-point scale formats are most commonly used, followed by five-point scales. Investigating four, five, six, and seven point scales, these authors also conclude that this is best practice. More specifically, they suggest to use seven point scales for populations high in cognitive ability, verbal skills and/or experience with questionnaires (such as students), but limit the number of scale points to five for the general population. For respondents that can handle many scale points, scales with more response categories are better than scales with a smaller number of response categories, not only because they lead to more information, but also because these scales reduce extreme responding. For respondents who have more difficulties with a large number of scale points, a scale with many (vs. fewer) response options induces confusion and misresponse (i.e., responding in the same direction to opposite items) (Weijters, Cabooter and Schillewaert 2010).

Odd or Even Number of Scale Points

Should response scales contain a midpoint or not? Some researchers believe that offering a midpoint decreases data quality because it may seduce respondents to simply select the midpoint without forming or reflecting on their true opinion (e.g., Klopfer and Madden 1980; Gilljam and Granberg 1993). Others advocate that respondents who hold a truly neutral or ambivalent opinion, should not be forced to take a stance (Schuman and Presser 1981).

Recent evidence suggests that providing a midpoint may actually increase data quality. Forcing respondents to choose sides may result in task-related distress and these negative

emotions may lead respondents to disproportionally select negative response options. In line with this explanation, several researchers found that offering a midpoint leads to a significant shift from negative response options to the midpoint (Velez and Ashworth 2007; O'Muircheartaigh, Krosnick and Helic 2000; Weijters, Cabooter and Schillewaert 2010). In addition, Weijters, Cabooter and Schillewaert (2010) report less extreme responding and less misresponse for scales with (vs without) a midpoint.

Unipolar or Bipolar Scale Format

A scale format can be unipolar or bipolar. In the unipolar scale format each item is represented by a single statement (e.g., I think this is a good ad), whereas the bipolar scale format uses two opposing statements for each item (e.g., I think this is a bad ad – I think this is a good ad). According to some researchers, the polarity choice largely depends on the construct under study. Attitudes (e.g., attitude toward the ad, attitude towards the brand, ad-evoked emotions, satisfaction) are considered to be bipolar and thus should be measured with a bipolar scale format, whereas behavior is considered to be a unipolar construct and thus should be measured with a unipolar scale format (Rossiter 2011). Other researchers argue that both unipolar and bipolar scale formats can be used to measure, for example, emotions (Bagozzi, Gopinath and Nyer 1999). In practice, several researchers appear to use both type of scale formats interchangeably for attitudes and behaviors (Cabooter et al. 2016).

Important to know though is that respondents activate a different knowledge base depending on whether unipolar or bipolar scale formats are used. For bipolar scale formats, two explicit statement poles are provided and respondents activate knowledge related to both the positive and the negative pole. Subsequently, they consider the response options on the left hand side as gradations of the negative pole (e.g., bad) and they consider the response options on the right hand side as gradations of the positive pole (e.g., good). Doing so, respondents treat the scale format as symmetrical. In contrast, for unipolar scale formats, only one explicit statement pole is provided (e.g., good). This makes respondents uncertain of the other pole (could be not good or bad), the consequence of which is that respondents give meaning to the majority of the response options on the basis of the explicit pole. Thus, for an item such as ‘I think this is a good ad’, respondents do not treat half, but more than half of the scale options as gradations of ‘good’ (Gannon and Ostrom 1996; Cabooter et al. 2016).

Due to this difference in the perception of symmetry, respondents show more agreement (i.e., more acquiescence) for bipolar as compared to unipolar scale formats. For example, Cabooter et al. (2016) found 42% agreement for the unipolar statement 'We are experiencing an improvement in our quality of life', but 66% agreement for its bipolar equivalent 'We are experiencing a deterioration in our quality of life' - 'We are experiencing an improvement in our quality of life'.

To conclude, bipolar scale formats have the advantage that they are easier to interpret for respondents and that the researchers also know for sure how the two poles are interpreted (bad – good), while unipolar scale formats leave the researchers unsure of how respondents interpreted the implicit pole (not good or bad). The disadvantage of bipolar scales are that they can inflate means and thus provide too optimistic a picture (Cabooter et al. 2016).

Scale Numbering

Researchers have the option to indicate response options with positive numbers only (e.g., from 1 to 7 for a 7-point scale) or with negative and positive numbers (e.g., -3 to +3 for a 7-point scale). In general, unipolar scale formats come with positive numbers whereas bipolar scale formats come with negative and positive numbers (Schwarz et al., 1991b; Krosnick and Fabrigar 1997). However, other combinations of polarity and numbering are also used in practice (Cabooter et al. 2016).

In what sense do respondents react differently to differently numbered scale formats? Scale formats with positive and negative numbers, as compared to scale formats with positive numbers only, come across as more intense, and as such widen the range of the scale (Judd and Harackiewicz 1980; Schwarz et al., 1991b). As people are more likely to avoid intense labels because they represent more extreme, exceptional positions (e.g., de Langhe et al. 2011), one would expect to find less extreme responding for scale formats with positive and negative numbers as compared to scale formats with positive numbers only.

Cabooter et al. (2016) find that scale numbering indeed lowers extreme responding for a bipolar scale, but not for a unipolar scale. The reason for the latter is that the use of positive and negative numbers for a unipolar scale not only widened the scale in respondents' minds, but it also made the unipolar scale symmetric (which has an opposite effect on extreme responding).

As a more general conclusion, Cabooter et al. (2016) found that scale numbering has a bigger impact for unipolar scale formats than for bipolar scale formats. Bipolar scales are clear and therefore respondents do not really need the numbers to give meaning to the response options, but unipolar scales are less clear and therefore respondents draw on the numbers. As a consequence, including positive and negative numbers (as compared to positive numbers only) makes a unipolar scale resemble a bipolar scale.

Scale Labels

Researchers need to choose (1) how many labels to use (e.g., verbal labels for all response options or endpoint labels only), as well as (2) which specific labels to use. Concerning the former, a scale format with fully labeled response options is easier to interpret both for respondents and researchers (Wildt and Mazis 1978), but increases respondents' tendency to agree and thus results in a positive bias (Tourangeau, Rips and Rasinski 2000; Weijters, Cabooter and Schillewaert 2010). If only a few options are labeled, these options are more salient and become more accessible to respondents, and consequently they are chosen more often (e.g., Krosnick and Fabrigar 1997; Posavac, Herzenstein and Sanbonmatsu 2003). As a consequence, a scale with verbal labels for the endpoints only increases extreme responding as compared with a fully labeled scale. At the same time it increases misreponse because the unlabeled categories are less clear (Weijters, Cabooter and Schillewaert 2010). Weijters, Cabooter and Schillewaert (2010) recommend to use fully labeled scales if researchers are interested in opinion measurement (i.e., means and percentages), but scale formats with endpoint labels only if researchers are interested in estimating relations between variables.

As far as the specific labels are concerned, research has shown that wider, more intense endlabels (i.e., "strongly agree" and "strongly disagree") leads respondents to avoid these labels and thus to select the intermediate response options more often than when narrower, less intense end labels are used (i.e., "agree" and "disagree") (Wyatt and Meyers 1987). Manipulating the amplifier of the endlabels (i.e., strongly (dis)agree vs. completely (dis)agree), Weijters, Geuens and Baumgartner (2013) found that familiarity mattered even more than intensity. Their results showed that the more respondents were familiar with an endlabel (i.e., the more often they used it in daily language), the more they endorsed the response option associated with this label. For example, US respondents are more familiar with the endlabels 'completely (dis)agree' than with the endlabels 'strongly (dis)agree'. As a consequence, they will select the endpoints less when the scale format uses 'strongly

(dis)agree' as endlabels as compared to when 'completely (dis)agree' is used. Especially when opinion measurement is important or when researchers want to make comparisons between, for example, different cultural groups, it is important to think carefully about which exact labels to use (see also Puntoni et al. in this issue).

ETHICAL ISSUES

In planning and conducting advertising research, a number of ethical considerations have to be taken into account. First of all, permissions may have to be obtained from participants, companies and institutions. Often, it is advisable to seek some form of *informed consent* from participants, that is, explain them what their participation entails and ask for their permission to be included in the study. This is particularly important when including participants from vulnerable target groups, such as children, in which case parental permission is called for. Informed consent is somewhat contradictory to the principle of disguising the true nature of the study during the briefing. The best thing to do may be to seek partly informed consent, explain the participants that they will be debriefed after the study, and disclose the true nature of the study and the manipulations used during a debriefing at the end of the study. In case existing real brands are used, it may be advisable to gain permission from the brand holding company for including their brands in the study. Finally, most research institutions or grant providers now have an ethical policy that requires permission for studies that include measurements with human subjects as, of course, all experimental advertising does. At all times, researchers should respect the privacy of participants and not collect, use or publish privacy-infringing data that are not absolutely necessary to answer the research questions.

An important ethical concern is to avoid questionable research practices (John, Loewenstein and Prelec, 2012). Fiedler and Schwarz (2015) tested the prevalence of ten of these practices in psychological research. Most of them are also relevant (and probably prevalent) in advertising research. In order of importance they are (Fiedler and Schwarz 2015, pp 17):

- Claiming to have predicted an unexpected outcome
- Selectively reporting studies regarding a specific finding that 'worked'
- Deciding whether to exclude data after looking at the impact of doing so regarding a specific finding
- Failing to report all dependent measures that are relevant for a finding

- Collecting more data after seeing whether results were significant in order to render non-significant results significant
- Failing to report all conditions that are relevant for a finding
- Rounding off *p* values (e.g., reporting a p value of .054 as .05)
- Stopping data collection after achieving the desired result concerning a specific finding
- Claiming that results are unaffected by demographic variables (e.g., gender) although one is actually unsure (or knows that they do)
- Falsifying data

The last three practices are very rare, at least as reported by researchers. However, researchers report doing the first five practices in between 30 and 50% of the time. Obviously, these practices should be avoided, although it is fair to say that the most prevalent ones are also the most ambiguous ones in that they do not always imply questionable practices. For instance, not always reporting all dependent measures can reflect justifiable behaviors (e.g., editorial requests). Focusing on studies that worked may reflect the reality that null results are not easily accepted for publication and thus discarded. Collecting more data after initial analyses may be an acceptable research practice when effect sizes appear to be smaller than expected. Excluding data after looking at the impact of doing so may simply reflect the correct behavior of finding out what, for instance, outliers do to one's conclusions. Discarding respondents on the basis of controls for inattentive or careless responding is equally good research practice and is actually always called for (Meade and Craig, 2012). Nevertheless, although some of the most prevalent questionable research practices may be justifiable, researchers are certainly called for full openness on their data collection and data analyses (e.g., sharing raw data set online, disclosing how many respondents they discarded and for what reasons, etc.).

CONCLUSION

Experimental advertising research requires well-developed, well-controlled and well-executed designs in which stimuli, samples, questionnaires and measures are methodologically correctly combined to provide a genuine and meaningful contribution to theory and practice. A wrong or inappropriate decision on any of the issues discussed above can completely invalidate the findings, and thus render the study meaningless, or at least its results questionable. There is no such thing as a perfect experimental study, and to several of the

issues discussed above there is no straightforward 'right' or 'wrong' answer, but it is of the utmost importance that researchers carefully consider every methodological step, and make the appropriate choices. As a general checklist for setting up a sound experimental advertising study could serve the following:

- Is there a meaningful theoretical and practical contribution?
- Are the (advertising) stimuli simple, realistic and well-controlled and embedded in a realistic advertising context?
- Is the sample of participants relevant for the study at hand, of appropriate size, and are sub-samples equivalent ?
- Does the order of the questions safeguard against any biases?
- Are a manipulation check and quality control questions taken in?
- Are construct valid, balanced scales used?
- Are 7-point scales used for student, MTurk or other highly educated or experienced samples, and 5-point scales for the general population?
- Where possible, are bipolar, fully-labeled scales used? (if the researcher is interested in estimating relations rather than opinion measuring, end point labels may be better)
- Can an unrelated, heterogeneous set of items be included to correct for response styles?
- Are the necessary permissions obtained?

-

REFERENCES

Aiken, Leona S., Stephen G. West, and Raymond R. Reno (1991), *Multiple Regression: Testing and Interpreting Interactions*, London: Sage.

Aleksandrovs, Leonids, Nathalie Dens, Peter Goos, and Patrick De Pelsmacker (2015), "Mixed-media modeling may help optimize campaign recognition and brand interest. How to apply the "mixture-amount modeling" method to cross-platform effectiveness measurement," *Journal of Advertising Research*, DOI: 10.2501/JAR-2015-025.

Anderson, Chris (2008), "The End of Theory," *Wired Magazine*, 16 (7), 16-07.

Angell, Robert, Matthew Gorton, Johannes Sauer, Paul Bottomley, and John White (2016), "Don't Distract Me When I'm Media Multitasking: Toward a Theory for Raising Advertising Recall and Recognition," *Journal of Advertising*, 45 (2), 198-210.

Atzmüller, Christiane, and Peter M. Steiner (2010), "Experimental Vignette Studies in Survey Research," *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* 6 (3), 128-38.

Bagozzi, Richard P., Mahesh Gopinath, and Prashanth U. Nyer (1999), "The Role of Emotions in Marketing," *Journal of the Academy of Marketing Science,* 27 (2), 184-206.

Barnette, J. Jackson (2000), "Effects of Stem and Likert Response Option Reversals on Survey Internal Consistency: If You Feel the Need, There Is a Better Alternative to Using Those Negatively Worded Stems," *Educational and Psychological Measurement,* 60, 361–370.

Baumgartner, Hans, and Christian Homburg (1996), "Applications of Structural Equation Modeling in Marketing and Consumer Research: A Review," *International Journal of Research in Marketing,* 13 (2), 139–161.

Baumgartner, Hans, and Jan-Benedict E.M. Steenkamp (2001), "Response Styles in Marketing Research: A Cross-National Investigation," *Journal of Marketing Research,* 38, 143-156.

Bearden, William O., and Richard G. Netemeyer (1999), *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*, Thousand Oaks, CA: Sage.

Bergkvist, Lars I., and John Rossiter (2007), "The Predictive Validity of Multiple-Item Versus Single-Item Measures of the Same Constructs," *Journal of Marketing Research,* 44 (2), 175-184.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. (2012), "Evaluating Online Labor Markets for Experimental Research: Amazon. com's Mechanical Turk," *Political Analysis,* 20 (3), 351-368.

Bollier, David, and Charles M. Firestone (2010), *The Promise and Peril of Big Data,* Washington, DC: Aspen Institute, Communications and Society Program.

Bottles, Kent, Edmon Begoli, and Brian Worley (2014), "Understanding the Pros and Cons of Big Data Analytics," *Physician Exec* 40 (4), 6-12.

Boyd, Danah, and Kate Crawford (2012), "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society,* 15 (5), 662-679.

Bruner, Gordon C. (2015), *Marketing scales handbook: multi-item measures for consumer insight research,* Forth Worth, Texas, USA: GCBII Production, LLC.

Bruner, Gorden C., Karen E. James, and Paul J. Hensel, (2001), *Marketing Scales Handbook, A Compilation of Multi Item Measures, Volume III*, Chicago: American Marketing Association.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling (2011), "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?," *Perspectives on Psychological Science*, 6 (1), 3–5.

Cabooter, Elke, Bert Weijters, Maggie Geuens, and Iris Vermeir (2016), "Scale format effects on response option interpretation and use," *Journal of Business Research* (forthcoming).

Calder, Bobby J., Lynn W. Phillips, and Alice M. Tybout (1981), "Designing Research for Application," *Journal of Consumer Research*, 8 (2), 197-207.

Campbell, Margaret C., and Kevin L. Keller (2003), "Brand Familiarity and Advertising Repetition Effects," *Journal of Consumer Research*, 30 (2), 292-304.

Carroll, James M., Michelle S. Yik, James A. Russell, and Lisa Feldman (1999), "On the psychometric principles of affect," *Review of General Psychology*, 3(1),14-22.

Catania Joseph A., Diane Binson, Jesse Canchola, Lance M. Pollak, and Walter Hauck (1996), "Effect of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior," *Public Opinion Quarterly*, 60 (3), 345-375.

Cauberghe, Verolien, and Patrick De Pelsmacker (2010), "Advergames: The Impact of Brand Prominence and Game Repetition on Brand Responses," *Journal of Advertising*, 39 (1), 5-18.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci (2014), "Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers," *Behavior Research Methods*, 46 (1), 112–130.

Churchill, Gilbert A. (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, 16 (1), 64–73.

Cox, Eli P., III (1980), "The Optimal Number of Response Alternatives For a Scale: A Review," *Journal of Marketing Research*, 17 (4), 407−422.

Dahlèn, Micael, and Frederik Lange (2005), "Advertising weak and strong brands: Who Gains?," *Psychology & Marketing*, 22 (6), 473-488.

Danaher, Peter J., and Dagger, Tracey S. (2013), "Comparing the relative effectiveness of advertising channels: A case study of a multimedia blitz campaign," *Journal of Marketing Research*, 50 (4), 517-534.

Davis Rachel E., Mick P. Couper, Nancy K. Janz, Cleopatra H. Caldwell, and Ken Resnicow (2010), "Interviewer Effects in Public Health Surveys", *Health Education Research,* 25 (1),14–26.

De Keyzer, Freya, Nathalie Dens, and Patrick De Pelsmacker (2015), "Is this for me? How consumers respond to personalized advertising on social network sites," *Journal of Interactive Advertising,* DOI: 10.1080/15252019.2015.1082450.

de Langhe, Bart, Stefano Puntoni, Daniel Fernandes, and Stijn M.J. van Osselaer (2011), "The Anchor Contraction Effect in International Marketing Research," *Journal of Marketing Research,* 48 (2), 366-80.

De Meulenaer, Sarah, Nathalie Dens, and Patrick De Pelsmacker (2015), "Have No Fear: How Individuals Differing in Uncertainty Avoidance, Anxiety and Chance Belief Process Health Risk Messages," *Journal of Advertising,* DOI: 10.1080/00913367.2015.1018465.

Dens, Nathalie, and Patrick De Pelsmacker (2010), "Attitudes toward the extension and parent brand in response to extension advertising," *Journal of Business Research,* 63 (11), 1237-1244.

Dens, Nathalie, and Patrick De Pelsmacker (2015), "Does poor fit always lead to negative evaluations? Extension advertising and perceived brand quality," *International Journal of Advertising,* DOI: 10.1080/02650487.2015.1057924.

Dens, Nathalie, Patrick De Pelsmacker, Nathalia Purnawirawan, and Marijke Wouters (2012), "Do You Like What You Recognize? The Effects of Brand Placement Prominence and Movie Plot Connection on Brand Attitude as Mediated by Recognition," *Journal of Advertising,* 41(3), 35-54.

Deetlefs, Jeanette, Mathew Chylinski, and Andreas Ortmann (2015), "MTurk 'Unscrubbed': Exploring the Good, the 'Super', and the Unreliable on Amazon's Mechanical Turk," *UNSW Business School Research Paper* 2015-20.

De Pelsmacker, Patrick, and Wim Janssens (2005), "Advertising for new and existing brands: the impact of media context and type of advertisement," *Journal of Marketing Communications,* 11(2), 113-128.

Duff, Brittany R-L., and Sela Sar (2015), "Seeing the Big Picture: Multitasking and Perceptual Processing Influences on Ad Recognition," *Journal of Advertising,* 44 (3), 173-184.

Elen, Maarten, Evelien d'Heer, Maggie Geuens, and Iris Vermeir (2013), "The influence of mood on attitude-behavior consistency," *Journal of Business Research,* 66(7), 917-923.

Eisend, Martin (2006), Two-sided Advertising: A Meta-analysis, *International Journal of Research in Marketing,* 23 (2), 187-198.

Eisend, Martin (2010), “Explaining the Joint Effect of Source Credibility and Negativity of Information in Two-Sided Messages,” *Psychology & Marketing*, 27 (11), 1032-1049.

Eisend, Martin, George R. Franke, and James H. Leigh (2016), "Reinquiries in Advertising Research," *Journal of Advertising*, 45 (1), 1-3.

Erdogan, B. Zafer (1999), "Celebrity endorsement: A literature review," *Journal of Marketing Management*, 15 (4), 291-314.

Evanschitzky, Heiner, Carsten Baumgarth, Raymon Hubbard, and J. Scott Armstrong (2007), “Replication Research’s Disturbing Trend,” *Journal of Business Research*, 60 (4), 411–15.

Fiedler, Klaus, and Norbert Schwarz (2015), “Questionable Research Practices Revisited,” *Social Psychological and Personality Science*, DOI: 10.1177/1948550615612150.

Fitzsimons, Gavan J. (2008), “Editorial: Death to dichotomizing,” *Journal of Consumer Research*, 35(1), 5–8.

Foos, Adrienne E., Kathleen Keeling, and Debbie Keeling (2016), "Redressing the Sleeper Effect: Evidence for the Favorable Persuasive Impact of Discounting Information Over Time in a Contemporary Advertising Context," *Journal of Advertising* 45 (1), 19-25.

Frison, Steffi, Marnik G. Dekimpe, Christophe Croux, and Peter De Maeyer (2014), ”Billboard and cinema advertising: Missed opportunity or spoiled arms?” *International Journal of Research in Marketing*, 31 (4), 425-433.

Gannon, Katherine M., and Thomas M. Ostrom (1996), “How Meaning is Given to Rating Scales: The Effects of Response Language on Category Activation,” *Journal of Experimental Social Psychology*, 32, 337-360.

Gibbons, Judith L., Jennifer A. Zellner, and David J. Rudek (1999), “Effects of Language and Meaningfulness on the Use of Extreme Response Styles By Spanish-English Bilinguals,” *Cross-Cultural Research*, 33 (4), 369–381.

Gilljam, Mikael, and Donald Granberg (1993), “Should We Take Don't Know For an Answer?”, *Public Opinion Quarterly*, 57 (3), 348−357.

Goedertier Frank, Niraj Dawar, Maggie Geuens, and Bert Weijters (2015), “The Effect of Brand Typicality on Distant Novel Extension Acceptance: How Risk-Reduction Advantages Help Overcome a Lack Of Perceived Category Fit,” *Journal of Business Research*, 68 (1), 157-165.

Greenleaf, Eric A. (1992), “Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles,” *Journal of Marketing Research*, 29 (2), 176-188.

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

Huyghe Elke, Julie Verstraeten, Maggie Geuens, and Anneleen Van Kerckhove (2016), "Clicks as a healthy alternative to bricks: How online grocery shopping reduces vice purchases," *Journal of Marketing Research* (forthcoming)

Iacobucci, Dawn, Steven S. Posavac, Frank R. Kardes, Matthew J. Schneider, and Deidre L. Popovich (2015), "The Median Split: Robust, Refined, and Revived," *Journal of Consumer Psychology, 25* (4), 690-704.

Irwin, Julie R., and Gary H. McClelland (2003), "Negative Consequences of Dichotomizing Continuous Predictor Variables," *Journal of Marketing Research*, 40 (3), 366–371.

James, William L., and Brenda S. Sonner (2001), "Just Say No to Traditional Student Samples," *Journal of Advertising Research*, 41 (5), 63-71.

John, Leslie K., George Loewenstein, and Drazen Prelec (2012), "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling," *Psychological Science*, 23 (5), 524-532.

Jordan, Lawrence A., Alfred C. Marcus, and Leo G. Reeder (1980), "Response Styles in Telephone and Household Interviewing: A Field Experiment," *Public Opinion Quarterly*, 44 (2), 210–222.

Judd, Charles M., and Judith M. Harackiewicz (1980), "Contrast Effects in Attitude Judgment: An Examination of the Accentuation Hypotheses," *Journal of Personality and Social Psychology*, 38 (3), 569-578.

Kerr, Gayle, Don E. Schultz, and Ian Lings (2016), "Someone Should Do Something": Replication and an Agenda for Collective Action." *Journal of Advertising*, 45 (1), 4-12.

Kirk, Roger G (1982), *Experimental Design: Procedures For the Behavioral Sciences* (2nd Ed.), Belmont, California: Brooks/Cole Publishing Company.

Klopfer, Frederick J., and Theodore M. Madden (1980), "The Middlemost Choice on Attitude Items: Ambivalence, Neutrality or Uncertainty," *Personality and Social Psychology Bulletin*, 6 (1), 97−101.

Knowles, Eric S., and Christopher A. Condon (1999), "Why people say ''yes'': A dual-process theory of acquiescence," *Journal of Personality and Social Psychology*, 77 (2), 379–386.

Krosnick, Jon A., and Leandre R. Fabrigar (1997), "Designing Rating Scales for Effective Measurement in Surveys," in *Survey Measurement and Process Quality*, Lyberg, Lars E., Paul P. Biemer, Martin Collins, Edith D. De Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin, eds., New York: Wiley, 141-164.

Krumpal, Ivar (2013), “Determinants of Social Desirability Bias in Sensitive Surveys: A Literature Review,” *Quality & Quantity*, 47, 2025–2047.

Kühnen, Ulrich (2010), "Manipulation checks as manipulation: Another look at the ease-of-retrieval heuristic," *Personality and Social Psychology Bulletin*, 36 (1), 47-58.

Kulkarni, Atul A., and Hong Yuan (2015), "Effect of Ad-Irrelevant Distance Cues on Persuasiveness of Message Framing," *Journal of Advertising*, 44 (3), 254-263.

Laczniak, Russell N. (2015), “The Journal of Advertising and the Development of Advertising Theory: Reflections and Directions for Future Research,” *Journal of Advertising*, 44 (4), 429-433.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014), "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, 343 (6176), 1203-1205.

Lenth, Russell V. (2001), “Some Practical Guidelines for Effective Sample Size Determination,” *The American Statistician*, 55, 187-193.

Mason, Winter, and Siddharth Suri (2012), “Conducting Behavioral Research on Amazon’s Mechanical Turk,” *Behavior Research Methods*, 44 (1), 1–23.

Maxwell, Scott E., and Harold D. Delaney (1993), “Bivariate Median Splits and Spurious Statistical Significance,” *Psychological Bulletin*, 113 (1), 181-190.

McClelland, Gary H., John G. Lynch Jr., Julie R. Irwin, Stephen A. Spiller, and Gavan J. Fitzsimons (2015), “Median splits, Type II errors, and false–positive consumer psychology: Don't fight the power,” *Journal of Consumer Psychology*, DOI: 10.1016/j.jcps.2015.05.006.

Meade, Adam W., and S. Bartholomew Craig (2012), “Identifying Careless Responses in Survey Data,” *Psychological Methods*, DOI: 10.1037/a0028085.

Menon, Rama (1993), “Statistical significance testing should be discontinued in mathematics education research,” *Mathematics Education Research Journal*, 5 (1), 4-18.

Motl, Robert W., and Christine DiStefano (2002), “Longitudinal invariance of self-esteem and method effects associated with negatively worded items,” *Structural Equation Modeling*, *9*(4), 562–578.

Naik, Prasad A. and Kalyan Raman (2003), “Understanding the impact of synergy in multimedia communications,” *Journal of Marketing Research*, 40 (4), 375-388.

Nunnally, Jim C. (1978), *Psychometric Theory*, 2d ed., New York: McGraw-Hill.

O'Muircheartaigh, Colm A., Jon A. Krosnick, and Armin Helic (2000), “Middle alternatives, acquiescence, and the quality of questionnaire data,” https://www.researchgate.net/profile/Colm_OMuircheartaigh/publication/5091207_Middl

e_Alternatives_Acquiescence_and_the_Quality_of_Questionnaire_Data/links/542971020cf238c6ea7f430c.pdf, retrieved, May 11, 2016.

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko (2009), "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power," *Journal of Experimental Social Psychology*, 45 (4), 867-72.

Orne, Martin T. (1962), "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications," *American Psychologist*, 17, 776-783.

Pany, Kurt, and Philip M.J. Reckers (1987), "Within- vs. Between-Subjects Experimental Designs: A Study of Demand Effect," Auditing: A Journal of Practice & Theory, 7(1), 39-53.

Paolacci, Gabriele, and Jesse Chandler (2014), "Inside the Turk: Understanding Mechanical Turk as a Participant Pool," *Current Directions in Psychological Science*, 23 (3), 184–188.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis (2010), "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, 5 (5), 411–419.

Park, Jung Hwan, Olesya Venger, Doo Yeon Park, and Leonard Reid (2015), "Replication in Advertising Research, 1980–2012: A Longitudinal Analysis of Leading Advertising Journals," *Journal of Current Issues and Research in Advertising*, 36 (2), 115–35.

Paulhus, Delroy L. (1991), "Measurement and Control of Response Bias," in *Measures of Personality and Social Psychological Attitudes*, John P. Robinson, Phillip R. Shaver, and Lawrence S. Wright, eds., San Diego: Academic Press, 17-59.

Paulhus, Delroy L. (2003), "Self-presentation measurement," In: *Encyclopedia of Psychological Assessment*, Fernandez-Ballesteros, R. (ed.), Thousand Oaks: Sage, 858–860.

Peter, J. Paul (1981), "Construct validity: a review of basic issues and marketing practices," *Journal of Marketing Research*, 133-145.

Pham, Michel Tuam, Maggie Geuens, and Patrick De Pelsmacker (2013), "The influence of ad-evoked feelings on brand evaluations: Empirical generalizations from consumer responses to more than 1000 TV commercials," *International Journal of Research in Marketing*, 30 (4), 383-394.

Podsakoff, Philip M., Scott B. MacKenzie, Jeong-Yeon Lee, and Nathan P. Podsakoff (2003), "Common method biases in behavioral research: a critical review of the literature and recommended remedies," *Journal of Applied Psychology*, 88(5), 879.

Posavac, Steven S., Michal Herzenstein, and David M. Sanbonmatsu (2003), "The Role of Decision Importance and the Salience of Alternatives in Determining the Consistency between Consumer's Attitudes and Decisions," *Marketing Letters,* 14, 47-57.

Quilty, Lena C., Jonathan M. Oakman, and Evan Risko (2006), "Correlates of the Rosenberg self-esteem scale method effects," *Structural Equation Modeling,* 13(1), 99–117.

Rajabi, Mahdi, Nathalie Dens, Patrick De Pelsmacker, and Peter Goos (2015), "Consumer responses to different degrees of advertising adaptation: the moderating role of national openness to foreign markets," *International Journal of Advertising,* DOI: 10.1080/02650487.2015.1110949.

Rand, David G. (2012), "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments," *Journal of Theoretical Biology,* 299, 172–179.

Reynar, Angela, Jodi Phillips, Simona Heumann (2010), "New technologies drive CPG media mix optimization," *Journal of Advertising Research,* 50 (4), 416-427.

Richter, Michael (2010), "Pay attention to your manipulation checks! Reward impact on cardiac reactivity is moderated by task context," *Biological Psychology* 84 (2), 279-289.

Rosnow, Ralph Leon, and Robert Rosenthal (1997), *People studying people: artifacts and ethics in behavioral research,* New York: Freeman.

Rossiter, John R. (2002), "The C-OAR-SE Procedure for Scale Development in Marketing," *International Journal of Research in Marketing*, 19 (December), 305–335.

Rossiter, John R. (2011), *Measurement for the Social Sciences,* New York: Springer.

Royne-Stafford, Marla B. (2016), "Research and Publishing in the Journal of Advertising: Making Theory Relevant," *Journal of Advertising,* 45 (2), 269-273.

Rucker, Derek D., Blakeley B. McShane, and Kristopher J. Preacher (2015), "A researcher's guide to regression, discretization, and median splits of continuous variables," *Journal of Consumer Psychology,* 25 (4), 666–678.

Schneider, Lars-Peter, and T. Bettina Cornwell (2005), "Cashing in on Crashes Via Brand Placement in Computer Games. The Effects of Experience and Flow on Memory," *International Journal of Advertising,* 24 (3), 321-343.

Schmidt, Susanne, and Martin Eisend (2015), "Advertising Repetition – A Meta-analysis on Effective Frequency in Advertising," *Journal of Advertising*, 44 (4), 415-428.

Schoenberg, Nancy E., and Hege Ravdal (2000), "Using Vignettes in Awareness and Attitudinal Research," *International Journal of Social Research Methodology,* 3(1), 63-74.

Schultze, Ulrike, and Julie Rennecker (2007), "Refraining Online Games," in: *Virtuality and Virtualization*: Springer, 335-51.

Schuman, Howard, and Stanley Presser (1981), *Questions and answers in attitude surveys: Experiments in question form, wording, & content*, New York: Academic.

Schwarz, Norbert, Herbert Bless, Fritz Strack, Gisela Klumpp, Helga Rittenauer-Schatka, and Annette Simons (1991a), "Ease of retrieval as information: Another look at the availability heuristic," *Journal of Personality and Social psychology*, 61(2), 195.

Schwarz, Norbert, Bärbel Knäuper, Hans-J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark (1991b), "Rating scales: Numeric values may change the meaning of scale labels," *Public Opinion Quarterly*, 55(4), 570−582.

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

Shaver, James P. (1992), *What statistical significance testing is, and what it is not*, Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Spencer, Steven J., Mark P. Zanna, and Geoffrey T. Fong (2005), "Establishing a Causal Chain: Why Experiments Are Often More Effective than Mediational Analysis in Examining Psychological Processes," *Journal of Personality and Social Psychology*, 89 (6), 845–51.

Spiller, Stephen, Gavan J. Fitzsimons, John G. Lynch, and Gary H. McClelland (2013), "Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression," *Journal of Marketing Research*, 50(2), 277–288.

Steenkamp, Jan-Benedict E.M., Martijn G. De Jong, and Hans Baumgartner (2010), "Socially desirable response tendencies in survey research," *Journal of Marketing Research*, 47 (2), 199-214.

Swain, Scott D., Danny Weathers, and Ronald W. Niedrich (2008), "Assessing Three Sources of Misresponse to Reversed Likert Items," *Journal of Marketing Research*, 45 (February), 116-31.

Taylor, Kevin M., and James A. Shepperd (1996), "Probing suspicion among participants in deception research," *American Psychologist*, 51(8), 886-887.

Tessitore Tina, and Maggie Geuens (2013), "PP for "Product Placement" or "Puzzled Public"? The Effectiveness of Symbols as Warnings of Product Placement and the Moderating Role of Brand Recall", *International Journal of Advertising*, 32(3), 419-442.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski (2000), *The Psychology of Survey Response*, New York: Cambridge University Press.

Uncles, Mark D., and Simon Kwok (2013), "Designing research with in-built differentiated replication," *Journal of Business Research*, 66 (9), 1398-1405.

Vandecasteele Bert, and Maggie Geuens (2010), "Motivated consumer innovativeness: concept and measurement," *International Journal of Research in Marketing*, 27(4), 308-318.

Van Kerckhove Anneleen, Maggie Geuens, and Iris Vermeir (2015), "The Influence of Looking Down Versus Up as a Learned Distance Cue on Level of Construal," *Journal of Consumer Research,* 41(6), 1358-1371.

Velez, Pauline, and Steven D. Ashworth (2007), "The impact of item readability on the endorsement of the midpoint response in surveys," *Survey Research Method*, 1(2), 69−74.

Verbeke, Willem, Paul Farris, and Roy Thurik (1998), "Consumer response to the preferred brand out-of-stock situation," *European Journal of Marketing*, 32 (11/12), 1008-1028.

Verberckmoes, Shana, Karolien Poels, Nathalie Dens, Laura Herrewijn, and Patrick De Pelsmacker (2016), "The Appropriateness of in-game advertising in fantasy games," *Proceedings of the 2016 EMAC conference*, Oslo: Norwegian Business School.

Weijters, Bert, and Hans Baumgartner (2012), "Misresponse to reversed and negated items in surveys: A review," *Journal of Marketing Research, 49*(5), 737-747.

Weijters, Bert, Hans Baumgartner, and Maggie Geuens (2017), "The Calibrated Sigma Method: An Efficient Remedy for Between-Group Differences in Response Category Use on Likert Scales," *International Journal of Research in Marketing*, 34 (1), forthcoming.

Weijters, Bert, Elke Cabooter, and Niels Schillewaert (2010), "The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels," *International Journal of Research in Marketing*, 27(3), 236-247.

Weijters, Bert, Maggie Geuens, and Hans Baumgartner (2013), "The effect of familiarity with the response category labels on item response to Likert scales," *Journal of Consumer Research,* 40(2), 368-381.

Weijters, Bert., Niels Schillewaert, and Maggie Geuens (2008), "Assessing Response Styles across Modes of Data Collection," *Journal of the Academy of Marketing Science*, 36, 409-422.

Wildt, Albert R., and Michael B. Mazis (1978), "Determinants of Scale Response: Label versus Position," *Journal of Marketing Research,* 15(2), 261-267.

Williams, Larry J., Joseph A. Cote, and Ronald M. Buckley (1989), "Lack of method variance in self-reported affect and perceptions at work: Reality or artifact?" *Journal of Applied Psychology*, 74(3), 462-468.

www.stat.ubc.ca/~rollin/stats/ssize/n2.html, accessed 21 March 2016.

Wyatt, Randall C., and Lawrence S. Meyers (1987), "Psychometric Properties of 4 5-Point Likert-Type Response Scales," *Educational and Psychological Measurement*, 47 (March), 27– 35.

Zaichkowsky, Judith L. (1994), "The Personal Involvement Inventory: Reduction, Revision, and Application to Advertising, " *Journal of Advertising*, 23 (4), 59-70.

TABLE 1

Sequence of measures in an experimental advertising study

Introduction or briefing
Manipulation
Dependent variables
Quality control
Mediating and moderating variables
Potential confounds – filler items
Manipulation check
Socio demographics
Suspicion probe
Debriefing