A New Method of Measuring Online Media Advertising Effectiveness: Prospective Meta-Analysis in Marketing

by

Gui Liberali

Glen Urban

Benedict Dellaert

Catherine Tucker

Yakov Bart

Stefan Stremersch

August 30, 2016

Gui Liberali is an Associate Professor of Marketing, Erasmus University, 3000 DR, T10-14, Rotterdam, The Netherlands, (+31)10-4082732, liberali@rsm.nl

Glen L. Urban is the David Austin Professor of Marketing, Sloan School of Management, Massachusetts Institute of Technology, 100 Main Street, Room E62-533, Cambridge, MA 02142, (617) 253-6615, glurban@mit.edu.

Benedict Dellaert is Professor of Marketing, Erasmus School of Economics, 3000 DR, H15-07, Rotterdam, The Netherlands, (+31) 10 - 408.13.53, Dellaert@ese.eur.nl

Catherine Tucker is the Sloan Distinguished Professor of Marketing, Sloan School of Management, Massachusetts Institute of Technology, 100 Main Street, Room E62-536, Cambridge, MA 02142, (617) 252-1499, cetucker@mit.edu.

Yakov Bart is an Assistant Professor of Marketing, Northeastern University - D'Amore-McKim School of Business, 202G Hayden Hall, Boston, Massachusetts, 857.3277373, y.bart@northeastern.edu

Stefan Stremersch is Chaired Professor of Marketing and Desiderius Erasmus Distinguished Chair of Economics at the Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, 3000DR, The Netherlands, and Professor of Marketing at IESE Business School, Universidad de Navarra, Barcelona, Spain. +31 10 408 8719, stremersch@ese.eur.nl

This work has benefitted from the comments of attendees of the 2016 Theory and Practice in Marketing conference (TPM), as well as the 2014 ECMI-AMA-EMAC Invitational Symposium on Marketing and Innovation, the 2014 and 2013 MIT Conference on Digital Business, and the special session on collaborative research in marketing at the 2011 Marketing Science Conference. We want to thank John Hauser for his contribution, input and suggestions, and to thank GM, MIT's Initiative on the Digital Economy, the Erasmus Center for Marketing and Innovation and INSEAD for partial financial support for this work. We want to thank the GM team - Andy Norton, Patricia Hawkins, Joyce Salisbury, Dan Roesch, David VanderVeen, Karen Ebben, Jonathan Owen, and Phil Keenan. We thank Patrick van Dijl, Ryan Ko, Qui Nguyen, Brian Baker, and Yasemin Goke for their research assistance support.

A New Method of Measuring Online Media Advertising Effectiveness: Prospective Meta-Analysis in Marketing

Abstract

The authors introduce a new method, prospective meta-analysis in marketing (PMM), to estimate consumer response to online advertising on a large and adaptive scale. They illustrate their approach in a field study in the U.S., China and the Netherlands, covering equivalent ad content on social media, online video, display banner, and search engines. The authors tested a conceptual framework based on attention and engagement using a technological solution that allow them to observe participants browsing and clicking activity in depth from their own residences, offices, or places of choice to use the tested media platforms, e.g., Facebook, Weibo, Google, Baidu and others. The authors show how consumers respond differently to the same ad depending on how distant they are from purchase, and uncover which channels are most appropriate to which user at different stages of the funnel. They also show how engagement and attention strengthen consumer response to advertising. The authors show how PMM produces *exploratory* findings, *confirmatory* findings, and *replications by* systematically organizing the incremental exploration of complex phenomena with cycles of discovery and validation.

Keywords: *online advertising, field experiments, multichannel marketing, purchase funnel, meta-analysis*

Measuring consumer response to advertising has been a challenge for firms for decades, and has become harder with the growth of online advertising.

Consider a global media manager for Chevy who needs to decide on online ad spending for video, social media, and online search for such different markets as the USA, China, and the Netherlands. Ad effectiveness can vary by media type (e.g., consumers may react differently to social media sponsored posts compared with online search sponsored results). Ad effectiveness can vary by region (e.g., the dominant search engine in China is Baidu, not Google). Ad effectiveness can also vary according to the temporal distance separating consumers from purchase (consumers further away from purchase may rely more on social media than on online videos). Finally, studying ad effectiveness may require smaller samples in controlled experimental laboratory settings, but field experiments provide more realism (Lewis and Rao, 2015). Faced with such complexity, our manager would like to have precise, reliable measures of consumer responses to *each* media channel-country combination across consumer types.

This paper introduces a new method, prospective meta-analysis in marketing (PMM), offering a unique synergy between field experimentation and meta-analysis, to measure and estimate consumer response to online advertising on a large and adaptive scale. PMM is an adaptation of the medical sciences' prospective meta-analyses of clinical trials to the study of marketing phenomena. PMM systematically organizes the incremental exploration of complex phenomena with successive cycles of discovery and validation through planning and adaptation that produce *exploratory* findings, *confirmatory* findings, and *replications*. We apply PMM in a large-scale field study in the U.S., China and the Netherlands, covering several online media channels, and four auto segments. We measure consumer response to equivalent ad content across all countries and

channels in a natural setting, in close cooperation with GM's operations in the countries concerned.

Our application of prospective meta-analysis includes multiple unique features, as compared to prior studies on online media effectiveness. First, our study contributes to the ad effectiveness literature by being the first to examine funnel-specific effects of consumer response to advertising across various media channels for equivalent ad content. We show that consumers respond differently to the same ad depending on how distant they are from purchase, and we uncover which channels are most appropriate to which user at different stages of the funnel. These findings provide much-needed guidance to marketing managers looking to efficiently tailor their campaigns and budgets to the channel and to different consumer decision stages. Second, we build a novel consumer-level dataset by simultaneously studying consumer response to four media channels in three countries. This allowed us to compare consumer response to advertising on two dominant social media platforms: Facebook and Weibo, which no study has done before (Draganska, Hartmann, and Stanglein, 2014 looked at various channels, but they did not include social media). Third, we develop and deploy a technological solution that provides us with a unique compromise between realism and control: Respondents participated in our field study from their own residences, offices or places of choice to use media channels such as Facebook, and yet we were able to observe their browsing and clicking activity in depth while they were browsing our social media and search stimuli on their own devices. This allowed us to develop and empirically test a conceptual framework to study response to online advertising based on attention and engagement.

Our application of PMM to online media effectiveness yields several interesting substantive insights. We show that social media advertising is most effective at earlier stages of the funnel,

while online video is more effective closer to purchase date. We also demonstrate how engagement and attention strengthen consumer response to online video and social media advertising, but when both online video and social media are simultaneously used at high levels of engagement, a negative and statistically significant effect dampens consumer response. Finally, we find consumer response to online advertising in China to be much weaker than in the western countries in our sample.

This study has several significant implications for marketing practice. First, we help managers deciding on online advertising spending, by providing separate measures of consumer response to advertising in various countries and channels, including online video, social media, search, and display banners. Second, we provide managers with an informed and balanced evaluation of PMM as a new method. This application shows that PMM is a powerful method of choice when a firm experiments with new channels, new media or new markets, because it allows analysts to explore and develop initial assumptions through the meta-design and meta-analysis of a sequence of field trials, very much as a drug company explores and verifies the treatment potential of a new drug. This was precisely the interest of General Motors (GM), who was a strong ally for this research, as the company has been building its online media.

The prospective meta-analysis method can be applied to study other complex phenomena, such as choice behavior on shared economy platforms (such as AirBNB or Xiaozhu), or investment behavior on crowdfunding platforms (such as Kickstarter). Many research contexts could benefit from PMM's more reliable causal inference, control or inventory of contingency factors and large sample size, especially when researchers need to reliably and precisely estimate small effects.

The paper proceeds as follows. First, we discuss key issues related to measuring ad effectiveness in online media advertising, and propose a conceptual overview of the exposure-attention-engagement process of consumer response to media advertising that guides our analyses. Next, we describe the new prospective meta-analysis method. Finally, we describe how we apply PMM to address the online media budget allocation problem across multiple countries, how each study influenced the subsequent study, and how all datasets were pooled. The paper closes with a discussion on promising avenues for future media research, managerial implications and future developments, and applications of our method to other substantive contexts beyond advertising and other types of data such as laboratory and secondary data.

MEASURING MEDIA EFFECTIVENESS

We revise past research on measuring consumer response to advertising in two parts. We first take a close look at *what to measure* (e.g., how consumers interact with advertising), and then we briefly review prior literature on *how to measure* media effectiveness.

Consumer Interacting with Online Media: Exposure, Attention and Engagement

Online media channels provide marketing managers with much more information than was available in traditional broadcasting channels such as open TV and radio. In broadcasting, consumers could receive, not transmit, information. In online channels, cues about consumer reaction to advertising are readily available, such as clicks on Facebook posts. Because of this more fine-grained information about consumer response to advertising, marketing managers can now use a more in-depth perspective on how consumers process and react to advertising. The Expo-

sure-Attention-Engagement process, described next, provides some structure to help analysts frame and model consumer response to advertising.

The Exposure-Attention-Engagement Process

The traditional approach to estimate consumer response to advertising is based on estimating the probability of *exposure* given TV or radio advertising schedules. However, online channels such as Facebook give users the option to actively interact with the advertising copy by re-loading, sharing, posting, and clicking, and this affects consideration and purchase. For example, consider a consumer who is checking her Facebook page when a sponsored story about Chevy is shown. She may or not pay attention to the story: Past research has shown that most ads receive no more than a single eye fixation (Pieters and Wedel, 2012). If she did see the advertising, she may decide to engage with it, such as by clicking on the story to learn a bit more about it. Recent longitudinal research has shown that this kind of engagement is important in explaining variation in the effect of advertising on choice (Tuchman, Nair, and Gardete, 2015).

Thus, the experience of interacting with advertising can be generally described as a three-event process. First, users are exposed to ad copy. Second, the advertisement either attracts their attention or not. Third, if they did pay attention, they decide whether to engage with the advertisement by clicking on it, posting, sharing or some other type of channel-specific interaction. Figure 1 illustrates this Exposure-Attention-Engagement (E-A-E) process.

Insert Figure 1 about here

The effect of online advertising on consideration or purchase is then conditional on having been exposed, paid attention, and possibly having engaged with the online ad copy. Given this multi-event conditional process, it is not surprising that unconditional media effects in some channels such as online display advertising can be as small as a fraction of a percentage (Urban

et al. 2014). Often advertising can affect consideration without necessarily making consumers engage with the advertisement, hence the bypass on Figure 1. For example a fan of electric cars may become informed about a new electric vehicle just launched by her preferred brand via a banner display. Upon seeing the banner ad, she may decide to visit the website or the dealer instead of clicking and engaging with the ad.

Temporal Distance to Purchase

Though the advertising literature consistently shows that consumers go through different stages of deliberation and information processing while making decisions, there is little agreement regarding the actual stages (Vakratsas and Ambler, 1999). For example, Abhishek, Fader and Hosanagar (2015) use a hidden Markov-model to generate four states (disengaged, active, engaged and conversion); Liberali, Urban, and Hauser (2013) used a continuous-time Markov model with three states (not-consider, consider, and purchase). Bleier and Eisenbeiss (2015) use information, consideration, pre-purchase, and purchase. Several researchers have used the term 'funnel' to collectively describe the states from a first exposure to actual online purchase (e.g., Li and Kannan, 2014; Hoban and Bucklin, 2015; Hu and Damangir, 2014, Abhishek, Fader and Hosanagar, 2015). Regardless of specific states and labels used, the *temporal distance to purchase* matters when measuring media effectiveness because preferences are revised and vary over time as consumers acquire information about products (Dyzabura, 2015), and response to online advertising can depend on how narrowly or broadly consumer preferences are construed (Lambrecht and Tucker 2013).

Methods to Measure Media Effectiveness

Ultimately, behavioral outcomes such as consideration and purchase are the measures of highest interest to managers, but these outcomes are difficult to measure (Hongshuang and Kan-

nan, 2014). Several types of studies are commonly used to estimate consumer response to advertising, notably laboratory studies, secondary-data studies, meta-analysis of secondary data studies, and field experiments. Laboratory studies are studies performed in highly controlled environments with relatively small samples of respondents, and often focus on uncovering the underlying psychological processes that drive consumer response to online advertisement. Secondary data are often studies through econometric analysis of observational data on advertising scheduling and spending (e.g., Vakratsas et al., 2014 and Naik and Peters, 2009). Field experiments tend to exploit randomized variation in exposure to advertising in the natural context of consumers as possible or an environment that is as close to that as possible (e.g., Draganska, Hartmann, and Stanglein, 2014), often focusing on a single advertising channel or technique (e.g., Lambrecht and Tucker 2013). Meta-analyses combine results from several studies in an effort to obtain greater statistical power and improve estimates of the size of the effect of interest.

Table 1 compares these four types of studies and our proposed method along four dimensions: the degree to which the studies are designed to claim causality, how much adaptivity a method gives to the researcher to incorporate new insights into the design of the subsequent studies, the level of data aggregation (individual-level data versus aggregate data), and how the studies handle consumer attention to advertisement (i.e., whether consumers are forced to pay attention to the advertisement or not, and whether consumer attention is observed or not), The right-hand column of Table 1 provides some information about the proposed method, PMM. More details on PMM are provided in the next section.

Insert Table 1 about here

Table 1 shows that the types of studies vary widely with respect to the four comparison criteria, namely causality, adaptivity, aggregation, and attention. First, causality is most easily tested

in laboratory studies and field experiments because the researcher is able to manipulate a specific condition per study while keeping other conditions and confounding factors fixed, or under strict control. Second, laboratory studies and field experiments follow a fixed protocol that establishes sample size, treatments, controls, and variables of interest. In contrast, secondary-data methods and meta-analyses rely on existing data sources often generated for other purposes hence the concept of a research-driven data-collection protocol is not directly applicable. Third, while laboratory studies and field experiments are based on individual-level data, the other types of studies often use aggregated data (e.g., sales data) generated for non-research purposes. Fourth, typical laboratory studies tend to ask consumers to look at specific advertisement copies, which to a large extent 'forces' respondents to pay attention. In such studies it is possible to observe whether and how much attention consumers paid to the stimuli. In contrast, secondary-data studies and meta-analysis use data that reflects consumers' (non-forced) decision to pay attention (or not) to advertisement in their natural context as part of their daily lives. This means that the researcher using secondary data or meta-analysis is rarely able to identify which consumers paid attention to the advertisement and which did not. Finally, in a typical field experiment, subjects are stimulated to pay attention to the advertisement, but typically the researcher cannot force or observe compliance.

In contrast to these four types of studies, PMM is appropriate for uncovering causal relationships, because it is based on a collection of successive and tightly coordinated field-experiment studies at the individual-consumer level. PMM allows the researcher to monitor consumer attention to advertisement because it uses individual-level data. Finally, PMM provides flexibility to incorporate insights from one study into the design of the next study through an adaptive proto-

col that outlines the parameters of the research design of all studies, which are revised upon the completion of each individual study. In the next section we provide more details on PMM.

PROSPECTIVE META-ANALYSIS IN MARKETING (PMM)

The aim of our research is to build on the advantages of being able to draw causal inference that is inherent to randomized ad testing and laboratory studies, whilst also allowing for the broad scope of research designs that cover multiple channels and that is used in econometric testing. To do so we develop Prospective Meta-Analysis in Marketing (PMM). We developed PMM inspired by the long tradition of large-scale collaborative research in the medical sciences. To a large extent we adapt Cochrane Collaboration's Prospective Meta-analysis (Higgins and Green, 2011) to the context of marketing research. Prospective Meta-Analysis requires that the researchers specify hypotheses, intended analyses, and selection criteria prior to knowing the results of individual studies (Higgins and Green, 2011). It allows large groups of researchers to integrate their efforts. For example, in medicine a prospective meta-analysis of 14 clinical studies and 90,000 patients studied the cholesterol-lowering effects of statins (Baigent et al., 2005).

Prospective Meta-Analysis has the potential to become the method of choice for large-scale experimental research in marketing because it is based on a systematic organization of randomized controlled trials (RCT). RCTs are widely used in the medical sciences, social sciences, and in marketing (Hayes et al., 2012). Medical researchers design and run RCTs to study the effect of medical drugs and treatments on patient health, conditional on genetic make-up of each patient. We developed PMM so that marketing researchers can design and run RCTs to test the effects of different marketing instruments on consumer behavior. More specifically, we propose

that advertising researchers run media trials to study the effect of messages - such as a TV commercial when watching TV online - on consumer buying behavior, conditional on latent preferences of each consumer. Table 2 compares clinical trials of medical drugs (column 2) and trials of media (column 3) along several dimensions.

- Insert Table 2 about here -

In clinical trial of medical drugs, patients are subjected to a medical treatment with the expectation of an improvement in their health status. The improvement is affected by their genetic make-up and, to some extent, by contextual factors such as health care systems and socio-economic conditions. In randomized trials of online advertising, consumers are subjected to a message with the expectation of a change in buying attitudes. The change is affected by the latent preference structure of the individual consumer and contextual factors such as competition, consumer expectations, disposable income, and time pressure. In both cases early stopping is possible, but the aim is different. Early stopping in medical trials aims to improve patient well-being (e.g., by assigning the best treatment to placebo subjects once there is enough confidence in the performance of the medical drug). In contrast, early stopping in randomized controlled trials in marketing aims to obtain higher levels of statistical power where it is most needed (e.g., in cells with smaller effects).

Despite several similarities, measuring subject response to marketing instruments differs from measuring response to medical treatment on a key dimension – the objects are more numerous and created much more rapidly in marketing. More specifically, marketing instruments such as online banners are developed and launched much more rapidly than medical drugs and outnumber them by several orders of magnitude. For example, Kinch et al. (2014) researched Food and Drug Administration (FDA) records between 1827 and 2013 and found that in almost

190 years the FDA approved only 1,453 new molecular entities. Meanwhile, in 2012 a typical U.S. Internet user was exposed to more than 1,700 banner ads per month, corresponding to more than 5 trillion ads that year (Comscore, 2013). Even if one were able to tease apart the number of unique ads from the number of repeated exposures, advertising content would still be found to be developed and launched at much greater speeds than medical drugs.

Because new advertising messages and formats are created very rapidly at relatively low cost, studies of consumer response to advertising must be more adaptive and flexible than studies of patient response to medical treatment. Thus, when we developed PMM we allowed for more flexibility that in the original prospective medical studies.

A PMM project is defined by the six elements in Table 2 (intervention, unit of analysis, context, driver dependent variable and desired outcome) and by two adaptation rules – the *between-sample adaptation rule* and the *within-sample adaptation rule*. We now describe both rules in detail.

Planning and Between-Sample Adaptation in PMM

PMM allows for *between-sample* adaptation, in the sense that the design of a study is only finalized once the insights from the preceding study are obtained about which key variables are relevant regarding the object of the study. For example, after collecting data about a new social media channel, it may become clear that the effects are dependent on the purchase funnel stage. Thus the second study will be explicitly designed to account for better funnel measures or messages that appeal specifically to certain stages of the funnel. In the application section of this manuscript we provide much detail on how and which insights and adaptations we performed between samples and countries.

PMM studies can vary substantially depending on how extensively the focal object has been studied before, which impacts the trade-off between planning and between-sample adaptation. The design of PMM studies that are based on well-studied product categories and marketing variables (such as pricing of coffee or yogurt) tend to have a low level of adaptation between samples and a high degree of planning and optimization. This can be achieved by applying optimal research design methods such as Farley, Lehman, and Mahn (1998) and Cox and Reid, (2000, pg. 169). The design of PMM studies based on emerging and dynamic industries such as online media or shared economy platforms tend to have relatively high levels of adaptation, but lower levels of optimization.

Within-Sample Adaptation in PMM: Early Stopping

PMM allows for *within-sample* adaptation through a clearly documented early stopping rule for each study. Early stopping means that one of the conditions of a randomized controlled trial will not receive any more subjects once the stopping criterion is met. The stopping criterion must be clearly documented *before* any data is collected. Early stopping is a possible strategy in medical and marketing studies if planned and documented in advance. The goal of early stopping is different in both fields. In medical trials the goal is to minimize the number of patients that are not getting the best treatment and minimize patient exposure to treatments that turn out to be harmful. In marketing the goal is to allocate more sample to conditions in which the effects are known to be smaller, hence they need more statistical power and sample size. For details on power analysis and our early stopping procedures, please refer to the Appendix C.

Outcomes of a PMM Study

The various outcomes of a PMM study vastly differ in the extent to which they are found in the literature and in different studies in a PMM project. Therefore, we classify findings as exploratory, confirmatory, and replications, as shown in Table 3.

- Insert Table 3 about here -

An *exploratory finding* is a result that is novel to the literature but does not hold in most tested studies, or cannot be tested in most studies. Such a finding is an important element of the scientific discovery process as it uncovers new insights that need to be confirmed in further scaled-up studies. For example, it is reasonable to expect that consumers with very high levels of engagement with ads may exhibit particular behaviors of interest to managers, but such conditions are not easy to anticipate or test in field studies. *Confirmatory findings* are results that were observed in most tested samples, but there may be boundary conditions or other covariates that need to be further explored in future scaled-up studies. *Replications* are findings that were observed in all or most samples and were previously found in the literature. For example, past research has shown that T.V. advertising positively affects sales (Lodish et al., 1995)

APPLICATION: ANALYSIS OF ONLINE MEDIA ADVERTISING EFFECTS USING PMM

We now present our application of PMM to assess the effect of advertising in online media on consumers' consideration of automobiles in three different countries. We start by describing the overall research design, treatments, dependent variable, and control variables. We then present our model specification and estimation. Next, we describe the results of each of three studies within the PMM and indicate how between studies 1 and 2 the results informed the opti-

mal plan and design for the subsequent study. We then show the estimates across all three studies based on pooled data. Detailed data in study 3 allow us to provide managerial interpretations based on the *Exposure-Attention-Engagement* framework of response to advertising. Thus, the application illustrates how managers can use PMM to design marketing research to address large-scale questions.

Overall Research Design

This study was based on samples of consumers of automotive vehicles in the U.S. (5,184 consumers), China (921 consumers), and the Netherlands (796 consumers)[1]. The three studies benefited from collaboration with GM, who used its North American, Asian and European operations to provide us with operational support, market details, and advertising campaign information. GM funded 75% of the expenses of the project. The remaining 25% were split between the three universities participating in the consortium: A major university in the northeastern U.S., a major French business school with an established campus in Asia, and a major Dutch university. In each country, we used a panel of consumers from the same reputable sample provider that has worked with the auto industry in several countries for more than twenty years.

Respondents were screened for the category of automobiles they were interested in. Respondents whose interest was exclusively in those categories for which we had no stimuli were excluded. Respondents were also screened for minimal involvement with media channel, and the time window they plan to buy a car. The screening details are described in Appendix B.

This PMM study implemented between-sample adaptation after each study, and within-sample adaptation in the form of early-stopping only in the last study (discussed later in this section). The first study was done in the U.S., followed by China and the Netherlands. Each study was separated by several months, giving us enough time to use the findings of one study to adapt

the design of the subsequent study. The U.S. results informed our decision as to whether to study search advertising and display advertising in China. The findings from the U.S. and China informed our decisions to include additional measures of attention, engagement, and self-selection in the Dutch study, and the decision to statistically optimize sample allocation across treatment cells.

Though it is unusual in marketing to allow one study to inform another study to this degree rather than all studies being predetermined, as we discussed in the previous section this aspect of PMM represents a powerful benefit of the methodology for researchers and managers who are interested in gaining the most information possible conditional on expenditures.

Next, we present the results of the three studies in our application in chronological order. This allows us to clearly discuss the dynamic updating of the study planning in PMM.

Treatments

Respondents were randomly assigned to one of the treatments and control cells. Respondents assigned to the control cell were asked to read a newspaper article on aluminum. We considered that such a story, while being related to the automotive industry, would not affect the outcome variables of interest. Respondents assigned to the treatment cell received stimuli according to their focal category. The basis of the PMM design was a 3 (country) by 4 (online media) between-subjects design in which consumers were shown different online ads that were personalized for the category of car that they were interested in. In the U.S. we focused on the following categories: small (Sonic), compact (Cruze) and mid/full-size (Malibu). In Europe we focused on small (Peugeot 107), medium (Opel Corsa) and large (Opel Insignia). In China we focused on micro, small (Aveo), medium (Cruze), and mid/full-size (Malibu).

The three countries selected (U.S., China, and the Netherlands) all represented important markets for our industry partner, and differed strongly in the terms of consumer preferences and consumer online media behavior. This made the three countries particularly suitable for a first application of PMM. In particular, in the spirit of PMM, across the studies we optimized our data collection to obtain information on the most relevant online channels and most important parameters when moving from the first to the third study.

We developed a Chrome plug-in to allow this research to have novel levels of control and realism by seamlessly deploying advertising copies and providing measures of attention and engagement. Depending on their assigned treatments, respondents were invited to visit and browse a major automotive website (in the U.S.), use their preferred social media website (Facebook in the US and Netherlands, Weibo in China), watch a TV shown on YouTube (all countries), make an online search on Google (U.S. and Netherlands) or make an online search on Baidu (China). Table 4 presents a summary of which platforms were used in each media channel in each country.

- Insert Table 4 about here -

During their experience on these media channels our Chrome plug-in inserted the GM display advertisement among the several other ads that were organically shown to them as part of a normal interaction with the channel. We did not force respondents to pay attention to our ads or to engage with them. Respondents had no information regarding which was the experimental ads and which were the non-experimental ads shown on Facebook, Weibo, YouTube, beside Baidu and Google search results, and on the auto website. By design - because we did not artificially create engagement with the ad - some users did not see the treatment banner ad in the webpage,

and others saw it but did not pay attention. This reflects how consumers interact with online advertising in the real world.

Online video websites such as YouTube are often used to watch TV shows. Social media rely heavily on social network effects and are often used to relate to friends and relatives. Hence, we chose to focus on these two channels in all countries, i.e., U.S., China and Netherlands. The online video treatment in all countries was based on 30-second GM commercials inserted as commercial breaks of popular TV shows in the focal country. The insertions happened after 2 minutes of viewing. The shows are "House" in the U.S., "Flikken Maastricht" in the Dutch study and "Legend of Zhen Huan" in China. Both shows and commercials, illustrated in Figure 2, were hosted on YouTube.

- Insert Figure 2 about here -

The social media treatment in all countries consisted of GM-sponsored stories inserted on Facebook (in the US and Netherlands) and Weibo (in China). We mimic reality by measuring attention and engagement with advertising beyond what is typically done in field experiments and field studies. Whenever a consumer logged into one of the platforms monitored in this study (such as Facebook), our plug-in placed our social media treatment as a sponsored story on top of the list of updates on their personal social media site, as illustrated in Figure 2. We went to great lengths to make sure there were no noticeable differences between our treatments and real-world advertising in each platform, including updating our Google Chrome plug-in when one of the social media sites (Facebook or Weibo) updated its look and feel during the fieldwork (analysis shown no significant differences before and after the change). We believe our design obtained as realistic estimates as possible.

In the first and second studies, we also used the Chrome plug-in to expose consumers to advertising while searching on Google (U.S.) and Baidu (China), and browsing on a major vehicle website in the US. For example, our plug-in positioned our GM advertising as the top sponsored link on their set of paid search results, as illustrated in Figure 2. When a respondent accessed Google or Baidu our plug-in placed our social media treatment as a sponsored story on top of the list of updates on their personal social media site or as result of the search.

Measures and Model

The key dependent variable in our analyses is the change in a consumer's consideration of the advertised car before and after being exposed to the online advertising ("lift"). Consideration was measured before and after exposure to the stimuli using a ten-point scale. In addition consumers were asked to respond to stimuli evaluation questions, cultural questions, and demographics. We observed their browsing behavior such as the number of clicks made on the channel and time spent on the questionnaire and on stimuli. The full questionnaire is available in Web Appendix WA.

Let $Consideration_i$, be a dummy that takes on a value of 1 if the focal car was considered and 0 otherwise. Let $Lift\_Consideration_i$ be the difference in consideration of the focal car after and before consumer i was exposed to our treatment. Each consumer could select – but was not required to select – up to ten cars every time. Our model includes the four main treatments, $Online\ Video_i$, $SocialMedia_i$, $Banner_i$, and $Search_i$, and their two-way interactions.

Let $StdWeeklyUsageofSocialMedia_i$ be the standardized number of times per week consumer i uses his social media platform. Let $Std.WeeklyUsageofOnlineVideo_i$ be the number of hours per week consumer i watched online videos. Both are mean-centered and standardized.

Preferences are construed over time, which affects consumer response to advertising depending on how far away in time they are from their intended purchase date (Lambrecht and Tucker 2013). Thus, rather than using discrete states of the purchase funnel effect, we opt to include in our model the time left until conversion. Let $OneYeartoPurchase_i$, $TwoYearstoPurchase_i$, $ThreeYearstoPurchase_i$, and $FourYearstoPurchase_i$ be indicator variables that take on 1 if the time until the next purchase as reported by consumer *i* is, respectively, between 0 and 12 months, 13 and 24 months, 25 and 36 months, or 37 to 48 months; and 0 otherwise. Consumers that report time until next purchase greater than 48 months were screened out. These indicators allow us to assess whether advertising is more effective at earlier stages (if negative) or later stages (if positive) of the funnel. The model is as follows:

$$Lift\ Consideration_i = \propto + \beta_1 Online\ Video_i + \beta_2 SocialMedia_i + \beta_3 Search_i + \beta_4 Banner_i + \beta_5 Online\ Video_i\ x\ SocialMedia_i + \beta_6 Online\ Video_i\ x\ Search_i + \beta_7 Online\ Video_i\ x\ Banner_i + \beta_8\ Search_i\ x\ Social\ Media_i + \beta_9 SocialMedia_i\ x\ Banner_i + \beta_{10} Banner_i\ x\ Search_i + \beta_{11} Std.WeeklyUsageofSocialMedia_i + \beta_{12} Std.WeeklyUsageofOnlineVideo_i + \beta_{17} FourYears\ to\ Purchase_i + \beta_{18} ThreeYearsto\ Purchase_i + \beta_{18} TwoYears\ to\ Purchase_i + \beta_{18} OneYear\ to\ Purchase_i + \varepsilon_i \quad (1)$$

We investigate whether the four channels have greater effect on consideration earlier or later in the funnel by interacting them with $FourYearstoPurchase_i$, $ThreeYearstoPurchase_i$, and $TwoYearstoPurchase_i$, and $OneYeartoPurchase_i$, as shown on the next section.

Analysis and Results of Study 1: US

The first study assessed the effect of four types of online media – video, social media, search and banners. The OLS estimates from the US study are presented in Table 5. The first column shows the effects of the main treatment, its interactions, and indicators for funnel stages. The second column shows in detail how the effect of advertisement varies over funnel stages[2].

We found the main effect of online video on lift to be positive and statistically significant, in line with published research on television advertising (such as Lodish et al., 1995). The effect seems to become stronger as the expected date of purchase nears, providing support for the proposed funnel effect of time until purchase.

The social media main effect is marginally significant but the estimates of social media funnel effects over time, shown in the second column of Table 5, are strongest and statistically significant for consumers who are most distant from purchase. Consideration of automotive vehicles tends to be rather a long process. The funnel measures suggest social media is most effective at early stages of formation of the consideration set. This finding can inform media planners on which type of channel and content they should focus when developing their campaign.

- Insert Table 5 about here -

The effect of display banner ads is not statistically significant at $p<0.05$ so we are cautious in interpreting it. It could well be that the effect unconditional on attention is so small that it requires a very large sample to be accurately estimated.

Analysis and Results of Study 2: China

For comparability purposes, we opted to study online video and social media in China. These effect sizes were found in the U.S. to be large enough to be relatively easily detected, so there was a reasonable chance we would be able to find them in China as well. We expected eventual differences between U.S. and China estimates to be informative because Weibo, not Facebook, is the major social media platform in China.

The findings from the US study affected the design of the study in China in accordance with the PMM approach. Specifically, we excluded studying banner ads in China. The lack of significance on the estimates of banner display advertising in the US study and past literature

suggests that the effect size of banner ads is so small (if present) that sample sizes need to be prohibitively large to detect them, despite the fact we were dealing with estimates conditional on exposure. Therefore, we did not predict subsequent studies would find larger effect sizes for display banners.

It is worth noting that we decided to include sponsored search advertising in the China study, even though it was not significant in the US. Unlike display banner advertising, the effect sizes of search advertising tend to be relatively large in published literature (e.g., between 0.16, and 0.18 in Dinner, van Heerde, and Neslin, 2014). Additionally, the dominant search engine in China is Baidu, not Google. We were interested in learning more about the effect of sponsored search advertising conducted on a popular search engine that is not Google. Table 6 shows the OLS estimates of response to online advertising in the China study.

As in the U.S. sample, the main effects of online video advertising were found to be strongly statistically significant in China. Table 6 suggests a synergy between social media and online video advertising in China, that was not found in the U.S.. This effect could inform the development of integrated advertising campaigns across channels. As in the U.S., the effects of online video advertising become stronger at later stages of the funnel, which is shown by the interaction of time until purchase and online video.

- Insert Table 6 about here -

The statistically significant positive effect of the interaction of social media and the funnel suggests that the response to social media is more pronounced for consumers who are two years before the expected purchase date. We will return later to the funnel effects of social media when we pool data across countries.

Analysis and Results of Study 3: The Netherlands

After the China study, we had collected enough evidence about the effect of online video advertising, which was the strongest effect in studies 1 and 2, to have a clear expectation of online video effect sizes in the Netherlands, and the corresponding statistical power requirement in term of sample size. Thus, we decided to stop assigning respondents to the online video cell of the Dutch study after it had enough power to replicate the results of online video in the U.S. and China. To avoid stopping due to spurious significance (Simonsohn et al., 2011, Jennison and Turnbull, 1999, and FDA, 1988 and 2010), we monitored effect sizes, statistical power, and evolution of p-values. These details can be found in Appendix C.

The insights and experience with the design and findings of the US and China studies allowed us to increase the richness of our overall study by adapting the design of the sample in the Netherlands in three ways. First, given that in China, like in the U.S., we found no significant effects for search, we considered the non-significance of this effect to be well established and dropped the search treatment from the study in the Netherlands. Second, we collected additional controls measures for attention and engagement, which allowed us to develop and test the Exposure-Attention-Engagement framework, the results of which are discussed in detail later in this section. Third, we investigated a possible lingering concern in our data in greater detail - whether self-selection, due to panel respondents' voluntary acceptance of the Chrome plug-in needed for our study, could be a concern. To investigate this issue, we first performed an exploratory analysis on the U.S. data, in which we found no evidence that pointed to such problems. In the Dutch study we decided to collect more detailed information about non-compliers for more detailed inspection. (The details of this analysis are discussed in Appendix A.) Table 7 shows the OLS estimates of the Dutch study.

- Insert Table 7 about here -

The main effect of online video advertising corresponds with the strong effect found in the other two studies. However, the funnel effects of online video are less comparable with the other studies. This may be because the boundaries between online video and social media are less marked in the Netherlands. Dutch respondents may perceive the (Internet-based) streaming video advertising as being closer to the (Internet-based) Facebook advertising than did respondents in the U.S. and China. This is likely driven by greater adoption on Internet TV and cable TV in the Netherlands, which has one of the highest Internet penetration rate in the world (at 90% in 2015)[3]. As more people perceive the boundaries between Internet-based TV and Internet-based social media becoming blurred, the differences between the effects of online video advertising and social media advertising may disappear as well. Finally, as in the U.S. study, response to social media advertising seems to be stronger at earlier stages of the funnel.

Pooled Results – All Countries and Media

We now combine consumers' response across studies, and pool[4] data to inspect overall media effects after country-specific effects are accounted for. We include all media tested in the study..

The estimates of consumers' response to online advertising using the pooled data are shown in Table 8. The second column shows the estimates of the channel-specific main effects. The third column shows how these effects vary by country, reflecting the estimates based on the separate (country-specific) data shown in previous tables. Column 4 shows how the effects change over the funnel and time without weighting. Column 5 shows the estimates based on weighted pooled data[5].

- Insert Table 8 about here -

Pooling did not alter the estimates substantially when compared with results from each country separately: The magnitude, direction, and statistical significance of most estimates has not changed. For example, online video is strong in all countries, and social media is more pronounced in the US sample. As expected, the fixed-effect results suggest consumers in China and the Netherlands are less responsive to advertising compared with consumers in U.S.

The pooled results confirm the opposite and complementary effect of online-video advertising and social-media advertising. The last two columns of Table 8 suggest the strongest effects of online video are found in the later stages of the conversion funnel. The strongest effects of social media are likely to be in the early stages of the funnel. Weak effects of social media in year 2 appeared in the weighted pooling – it reflects the influence of the China sample (in the non-weighted pooling the China effects are weakened by the much larger U.S. sample). Banner advertising also seems to be more effective at early stages of the funnel. The managerial implications of such funnel-specific findings for campaign are discussed in the next section.

The E-A-E Model of Consumer Response to Online Advertising

We now turn to the conceptual Exposure-Attention-Engagement (E-A-E) process depicted in Figure 1. In study 3(the Dutch sample) we were able to include measures of attention and engagement. By conditioning on attention and engagement we show and measure how consumer response becomes increasingly stronger.

We condition on attention using post-measures of recall. More specifically, we use the dummy variable $attention_i$ that takes on a value of 1 if consumer i correctly recalled the car model she saw in the advertising, and 0 otherwise (in the social media cell this question was based on brand

recall because they were not shown a specific car model). This variable is based on a recall question asked soon after exposure to stimuli.

We condition on high levels of *engagement* with online advertising in social media using clickstream on the social media platform. More specifically, our Chrome plug-in allowed us to non-intrusively collect the number of clicks consumers made while they were browsing their experimentally-manipulated social media personal accounts. Thus, we define a dummy variable $engagement_i$ that takes on a value of 1 if consumer i in the social media cell is above the mean number of clicks in the sample on the website and 0 otherwise. This is a measure of high levels of engagement. This variable is left-censored because the sample is based on minimal level of clicks.

Given these two measures, we re-ran the analysis based on Equation 1 with two new variants. First, we condition the analysis on attention by restricting the sample to respondents that have $attention_i$ set to 1. Next, we condition the analysis on both attention and engagement by restricting the sample to respondents that have both $attention_i$ and $engagement_i$ set to 1. The results are shown in Table 9.

The first column of Table 9 replicates the original baseline estimates of the Dutch study without controlling for attention and engagement beyond the minimal screening criteria. The second column shows the estimates based only on consumers that paid attention to the stimuli, i.e., conditioning the data on $attention_i$. The third and last column shows the estimates given attention and engagement, i.e., further conditioning the data both on $attention_i$ and $engagement_i$.

- Insert Table 9 about here -

The comparison between the second and third columns shows that consumer response to online advertising and social media advertising is stronger when Dutch consumers are paying at-

tention to advertising – most main effects are stronger. The comparison between the third and fourth columns shows consumer response to online media advertising is stronger when consumers pay more attention and are more engaged - the estimates of social media advertising effects are significant in the third column. The negative interaction between social media and online video among consumers with high levels of engagement is an interesting finding that deserves further attention in future scale-up studies.

The unconditional effect of online advertising has been documented in past research but Table 9 shows consumer response specifically for consumers who have higher levels of attention and engagement with the media channel. This finding can be used to customize advertising efforts. Current targeting technology - varying from simple cookie-based behavioral targeting to optimization algorithms (e.g., Urban et al. 2014) - can easily allow for individual-level adaptive advertising that changes content or channel as respondents show increasing level of attention and engagement. Table 9 indicates how strong responses can be expected to be. For example, it shows the saturation effect resulting from the interaction of online video and social can decrease the effectiveness of a multi-channel campaign. Consequently, high-engagement campaigns have to be carefully designed to balance different media channels.

DISCUSSION

We start with a discussion of future research using PMM describe the managerial implication of our substantive findings. We close with a discussion on the limitations of PMM.

Discussion of Prospective Meta-Analysis and Future Applications

We proposed PMM as a framework for designing and running large-scale marketing research projects across multiple countries, studies and teams of researchers that relies both on planning and adaptation with a clear protocol and documentation detailing in advance key analyses and sampling procedures that will be performed, including early stopping. By applying PMM to the media effectiveness evaluation problem, we show how to design and run a PMM study in a comprehensive context because there is substantial variation across countries both on the supply side (e.g., social media websites in U.S. and China are very different) and on the demand side (consumers react differently to social media advertising in U.S., China and the Netherlands). Such variation requires large samples and geographically distributed teams, which increases field costs for academic studies that aim for high external validity. More importantly, this context provided various insights on how to use our method to address an inherently difficult problem that requires high levels of planning and adaptation.

Running a prospective meta-analysis project involves a great amount of tension and discussion until convergence is reached on decisions regarding funding sources, realism, inductivism, within-study adaptive sampling, and across-studies adaptation. The treatments should not be too close to existing practice nor too far away. It is a difficult trade-off. Treatments should not exactly mirror current practices in industry, otherwise findings are not likely to be applicable in other cases. However, treatments should not be so far removed from current practice that the findings are unlikely to be actionable. The optimal point tends to be such that treatments are based on the concepts underlying current practice. For example, rather than focusing on Facebook advertising, a study can focus on social media advertising in general, treating Facebook as just one (quite relevant) implementation and having other social media tools such as Weibo also in the pool of

treatments, preferably in countries where Facebook is not as prevalent as it is currently in the U.S.

Our within-study adaptation was very simple from a statistical point of view. More advanced methods are readily available to be used such as propensity score (Cohen, 1988), Bayesian predictive probabilities (FDA 2010 and Wang, Bradlow and George, 2014), classical Bayesian experimental designs (Berry 1991, and Chaloner, 1995), methods to reduce collinearity between design variables (Farley, Lehmann and Mann, 1998), and multi-armed bandits (Gittins, Glazebrook, and Weber, 2011; and Berry and Fristedt, 1985). There have been renewed calls for the use of multi-armed bandit models for early stopping and adaptive sampling in clinical trials (e.g., Press, 2009). Due to its ease of use and rapid closed-form updating, such solutions to bandit problems can be implemented either with optimal index-based policies such as the ones used in Urban et al. (2014) or – if only aggregated data is available - with heuristics such as Thompson sampling (Schwartz, 2013). It is also possible to use weighted regression on pooled data in between studies as detailed in Appendix B. Independent of the method used, our approach provides an overall framework that integrates all studies and allow for adaptation in research design and sampling strategies.

Regardless of the method chosen to perform adaptive sampling and early stopping, care must be taken when choosing the outcome variable to be used in the stopping rule because the scope is narrowed by the chosen outcome variable. For example, a researcher interested in the question, 'how does consumer exposure to warning labels on tobacco products affect consumer behavior?,' might use a within-study stopping rule focused on purchase likelihood, but that is not the optimal strategy for studying the effect on consumption likelihood.

Marketing researchers have been traditionally less open to early stopping of experiments than other fields such as the biomedical sciences. For example, the FDA not only accepts early stopping of clinical trials of medical treatments but also requires it to be planned and documented in advance, as a way to reduce the impact of poor treatments on patient welfare and safety. Biomedical researchers have had more than three decades of remarkable success with interim analysis in clinical trials, often with funding from the National Science Foundation and the National Institutes of Health (Montori, Deveraux, and Adhikari, 2005). Industry researchers also use interim analysis, under strict guidelines of the Food and Drug Administration (Jennison and Tunbull, 1999 and FDA, 2010).

As adaptive methods and stopping rules become more popular in marketing, we expect that the differences between within-sample adaptations (such as early-stopping) and across-sample adaptations (such as one study affecting the design of the next study) will become less visible and useful. The sample-view of the world that is dominant in marketing has its roots on the works of Ronald Fisher, who developed his statistical work on agricultural crops. In that substantive context, discrete and separate samples made sense because the outcome of a trial becomes known long before the next trial has been designed and started (Armitage, 1993). As marketing moves towards digital samples we may see p-values, effect sizes and stopping times being monitored and reported after each and every respondent (e.g., Simonsohn, and Simmons, 2014).

We provided a proof-of-concept in a framed field experiment that mimics reality in a natural setting (List, 2008), but prospective meta-analysis suits well different types of research design such as laboratory experiments, surveys, and full-fledged field experiments.

Implications of our Findings for Advertising Campaign Planning and Development

The global media manager for Chevy facing the online advertising spending decision described in the first section of this paper would find help in this research in several ways. First, we provide separate measures of consumer response to advertising in various channels, including online video, social media, search, and display banners. Second, because we take into account the effect of the temporal distance separating each consumer from purchase, we show how different channels affect consumers at different points in the funnel. For example, social media tends to affect consumers at earlier stages of consideration and online video affects consumers closer to purchase. Temporal distance makes causality hard to measure in longitudinal studies, especially in the online world. We assessed the distance-to-purchase for each respondent in our samples, and used fixed effects to obtain estimates of the differential effects of each channel along the purchase funnel. Third, we ran the study in geographically and culturally distant countries, namely, the U.S., China, and the Netherlands. In doing so we covered markets with different dominant players. This was important because Alphabet's Google search engine has almost a virtual monopoly in the search industry in western countries. By including China we were able to show similarities and differences on consumer response to advertising on a non-Google yet dominant platform, namely Baidu. For example, response to online video advertising is an order of magnitude weaker in China if compared with western countries in our sample.

Results in Table 8 illustrate how our method can be used to obtain comparable estimates across contexts and markets, and how these can be used in the planning of multi-channel advertising campaigns. For example, because social media advertising has an effect at early stages of consideration across all countries, advertising campaign content could be tailored towards early stages of consideration - when consumers tend to be searching for car models to add to the consideration set, rather than focusing on comparisons and purchase.

CONCLUSIONS

We addressed the problem of media budget allocation by measuring consumer response to online media advertising of automotive vehicles in three countries located in different continents with consumers at different stages of the consideration funnel and under various levels of attention and engagement. We confirmed the strong effects of online video advertising, and uncovered surprising funnel-specific effects of online video advertising and social media advertising. We also showed how engagement and attention strengthen consumer response to online video and social media advertising, but when both online video and social media are simultaneously used at high levels of engagement, a negative and statistically significant effect greatly dampens consumer response. Though search was not significant, the sponsored link directed consumers to the manufacturer's website rather than local dealers, who potentially may have had a different persuasive effect.

The outcomes of this PMM study differ in the extent to which they are found in the literature and in different samples. Table 10 summarizes our substantive results using the PMM typology of outcomes.

- Insert Table 10 about here -

We found that engagement with advertisement and attention to the advertising copy strengthen consumer response to advertising in two channels. We also found a negative interaction of response to video and social media at high levels of attention and engagement in the Dutch sample. However, both findings were only testable in one sample, so further research is needed to confirm them. We observed in most studies that response to social media advertising tends to be

more effective at early stages of consideration. Finally, the positive and strong effect of online-video advertising on consideration was confirmed in all our samples and was already expected.

Managerial Implications of PMM

PMM is not limited to academic research. Many large firms must decide how to allocate their advertising budgets across channels and countries, so their marketing research managers and agencies often struggle between ad-hoc early-stopping practices and rigid planning. Both can be very ineffective. Extreme and rigid planning wastes resources because of opportunity costs due to lessons that could have been learned between samples. Ad-hoc early stopping practices lead to p-hacking and spurious results. Managers of marketing research departments can use PMM to properly plan, document, and manage the trade-off between planning and adaptation. PMM gives the manager a systematic approach to balance the amount of planning and adaptation between samples, through cycles of discovery and validation.

PMM is appropriate to the study of complex marketing phenomena, well beyond consumer response to advertising. For example, PMM can be used both as a confirmatory tool to integrate estimates on widely studied topics as well as a way to systematically organize the incremental exploration of emerging and little studied consumer phenomena. Confirmatory studies using PMM could include pricing, sales and promotion of traditional product categories such as durable goods, appliances and fast-moving goods. Exploratory studies could include settings that change rapidly such as e-commerce and the sharing economy (as Airbnb and Uber).

Implications for Future Research of PMM

There are several limitations to our substantive findings that are worth noting. First, future studies could expand the depth of the findings by understanding the effect of media at the attribute level. During the China study, it became clear that some social media sites such as Weibo

combine features that are typical of social media (such as networking with friends) and others that are specific to micro-blogging (such as posting instant messages). Being able to separate the effect of each feature of media channel could be informative for advertisers as well as media firms.

Future studies could also expand the breadth of current findings in three ways. First, we do not attempt to provide optimal advertising allocation budgets on the basis of our results. Our results provide a basis for managerial action and new insights into how consumers respond to media. This provides a foundation for future work to build a media advertising budget optimization allocation model that uses these estimates as input for an optimization model such as that conducted by Danaher, Lee, and Kerbache (2010), which would then produce specific policies indicating how much to spend in each channel given costs and exposure rates in each specific market. Second, though we focus on many forms of new and emerging media, there are other forms of new media we do not consider, such as Instagram, Pinterest, earned media and mobile advertising. Third, we focus consumer response to media for a durable high-priced good. Therefore, these findings may not generalize for goods which have a less involved selection process. Fourth, we used a proxy for engagement based on the viewing time and total number of clicks. This was reasonable in this field study because respondents were invited to use the platform while our tailor-made Chrome plug-in monitored their activity. As ads become more sophisticated (e.g., interactive story-telling, apps and games), it is reasonable to expect that new measures of engagement will become available. Notwithstanding these limitations, we believe that our research provides substantively useful insights into how multiple media channels affect the consumer purchase decision process.

REFERENCES

Abhishek Vibhanshu, Fader Peter, Hosanagar Kartik (2015), "Media exposure through the funnel: A model of multi-stage attribution," *Working paper*, Heinz College, Carnegie Mellon University, Pittsburgh

Armitage, Pete (1993), "Interim Analyses in Clinical Trials," In *Multiple Comparisons, Selection, and Applications in Biometry*, ed. F.M.Hoppe, New York: Marcel Dekker, pp. 391-402.

Baigent C, Keech A, Kearney P M, Blackwell L, Buck G, Pollicino C, Kirby A, Sourjina T, Peto R, Collins R, Simes R. (2005), "Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins," *Lancet*, 366, 1267-1278

Bleier, Alexander and Eisenbeiss, Maik (2015), "Personalized Online Advertising Effectiveness: The Interplay of What, When, and Where," *Marketing Science,* 34(5):669-688.

Berry, Don (1991), "Experimental design for drug development: a Bayesian approach," *Journal of Biopharmaceutical Statistics*, 1:1, 81-101

Berry, Don and B. Fristedt (1985), *Bandit Problems. Sequential Allocation of Experiments*. Monographs on Statistics and Applied Probability. London/New York: Chapman and Hall.

Chaloner, Kathryn (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, Vol. 10, No. 3, pp. 273-304

Cohen, Jacob, (1988), *Statistical Power Analysis for the Behavioral Sciences*. New York: Psychology Press.

Comscore, (2013),*US Digital Future in Focus 2013*. Chicago: comScore Inc.

Cox, David and Reid, Nancy (2000), *The Theory of the Design of Experiments*. Monographs on Statistics and Applied Probability. New York: Chapman and Hall/CRC.

Danaher, Peter, Lee, Janghyuk and Kerbache, Laoucine (2010), "Optimal Internet Media Selection," *Marketing Science*, 29:2, 336-347

Dinner, Isaac Van Heerde, Haral and Neslin, Scott (2014), "Driving Online and Offline Sales: The Cross-Channel Effects of Traditional, Online Display, and Paid Search Advertising," *Journal of Marketing Research*, 51(5), 527-545.

Draganska, Michaela, Hartmann, Wesley and Stanglein, Gena (2014), "Internet Versus Television Advertising: A Brand-Building Comparison," *Journal of Marketing Research*, 51(5), 578-590.

Farley, John, Lehmann, Donald, and Mann. Lane (1998), "Designing the Next Study for Maximum Impact," *Journal of Marketing Research,* 35(4), 496-501.

FDA - Food and Drug Administration (1988), "Guideline for the Format and Content of the Clinical and Statistical Sections of New Drugs Applications," *FDA Tech. Rep.* (1988).

FDA - Food and Drug Administration (2010), "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials"

Gittins, John. Glazebrook, K. and Weber, Richard. (2011), *Multi-Armed Bandit Allocation Indices*. London: Wiley & Sons.

Greene, William (2011), *Econometric Analysis*. Upper Saddle River, New Jersey: Wiley.

Higgins Julian, Green Sally (2011). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0. The Cochrane Collaboration. Available from www.handbook.cochrane.org.

Hoban, Paul and Bucklin, Randolph (2015), "Effects of Internet Display Advertising in the Purchase Funnel: Model-Based Insights from a Randomized Field Experiment," *Journal of Marketing Research,* 52, 375-393.

Hongshuang, Li and Kannan, P.K. (2014), "Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment," *Journal of Marketing Research*, 51 40–56.

Jennison, Christopher and Turnbull, Bruce (1999), *Group Sequential Methods with Applications to Clinical Trials.* New York: Chapman

Kinch Michael, Haynesworth Austin, Kinch Sarah, Hoyer Denton (2014), "An overview of FDA-approved new molecular entities: 1827-2013," *Drug Discovery Today*, 19 (8):1033-9.

1. J. Simmons, L. Nelson,. U. Simonsohn, False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci*. **22** (2011).

Lambrecht Anja. and Tucker Catherine (2013), "When does retargeting work? Information specificity in online advertising," *Journal of Marketing Research*, 50(5):561–576.

Lewis, Randall and Rao, Justin (2015), "The Unfavorable Economics of Measuring the Returns to Advertising," *The Quarterly Journal of Economics* first published online July 6, 2015 doi:10.1093/qje/qjv023

Liberali, Guilherme, Urban, Glen. and Hauser, John. (2013), "Competitive Information, Trust, Brand Consideration and Sales: Two Field Experiments," *International Journal of Research in Marketing,* 30 101–113.

List, John (2008), "Introduction to field experiments in economics with applications to the economics of charity," *Experimental Economics*, 11:203-212.

Lodish, Leonard,Abraham, Magrid, Kalmenson, Stuart, Liverlsberger, Jeanne, Lubetkin, Beth, Richardson, Bruce, and Stevens, Mary (1995), "How TVAdvertising Works: A Meta-Analysis of 389 Real World Split Cable T V Advertising Experiments," *Journal of Marketing Research*, 32: 125-139

Montori, Victor, Devereaux, P., and Adhikari, et al.(2005), "Randomized Trials Stopped Early for Benefit: A Systematic Review," *JAMA*, 294

Naik, Prasad and Peters, Kay (2009), "A Hierarchical Marketing Communications Model of Online and Offline Media Synergies," *Journal of Interactive Marketing*, 23: 288–299

Pieter, Rik, and Wedel, Michel (2012),"AdGist: Ad Communication in a Single Eye-Fixation," *Marketing Science*, 31 (1): 59-73.

Press, William (2009),"Bandit Solutions Provide Unified Ethical Models for Randomized Clinical Trials and Comparative Effectiveness Research," *PNAS*, 106(52), 22387–22392.

Schwartz Eric (2013),"The attribute-based multi-armed bandit for adaptive marketing experiments*". Dissertation thesis*, University of Pennsylvania, Philadelphia.

Simmons, Joseph, Nelson, Leif and Simonsohn, Uri (2011),"False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22

Simonsohn, Uri, Nelson, Leif, and Simmons, Joseph, (2014),"P-Curve: A Key to the File-Drawer," *Journal of Experimental Psychology: General*, 143.

Tuchman,A., Nair,H., and Gardete Pedro. (2014),"An Empirical Analysis of Complementarities Between the Consumption of Goods and Advertisements, Stanford Working Paper Series.

Urban Glen, Liberali Guilherme, MacDonald Erin, Bordley Robert, Hauser John (2014),"Morphing Banner Advertising, Marketing Science, 33(1): 27-46.

Vakratsas , Demetrios, and Ambler, Tim (1999),"How Advertising Works: What Do We Really Know?," *Journal of Marketing,* 63 (1): 26-43

Vakratsas , Demetrios and Kolsarici , Ceren (2014),"*How DTCA Influences Prescription Pharmaceutical Markets,"* IN: Ding, Min; Eliashberg, Josh and Stremersch, Stefan (eds.), Innovation and Marketing in the Pharmaceutical Industry, International Series in Quantitative Marketing. New York: Springer-Verlag.

Wang, Pengyuan, Bradlow, Eric and George, Edward (2014),"Meta-Analyses Using Information Reweighting: An Application to Online Advertising," *Quantitative Marketing and Economics*, 12, 209 – 233.

FOOTNOTES

1 - The Dutch study included the Dutch-speaking Belgian Flanders, due to its deep linguistic affinity with the Netherlands. For ease of exposition, we refer to the area of this study simply as the Netherlands.

2 - All estimates in Tables 5, 6, and 7 are relative to subjects assigned to control group that are one year away from purchase, and conditional on the screening criteria described in Appendix B.

3 - https://en.wikipedia.org/wiki/Digital_television_in_the_Netherlands

4. The technical details of pooling and weighting procedures for all samples are outlined in Appendix B

5- All estimates in Tables 8 are relative to subjects assigned to the control group in the U.S. that are one year away from purchase, and are conditional on the screening criteria detailed in Appendix B.

TABLES

Table 1 – Methods to Study Media Effectiveness

	Laboratory Study	Secondary-Data Study	Meta-Analysis	Single Field Experiment	Our Method (PMM) Applied to Advertising
Causal Claims	Yes	No*	No	Yes	Yes
Adaptivity	Protocol	No Protocol	No Protocol	Protocol	Adaptive Protocol
Data Aggregation	Individual-level	Aggregated data	Aggregated or Individual data	Individual-level	Individual-level
Attention to Advertisement	Forced, always observed	Natural, often not observed	Natural, often not observed	Stimulated, often not observed	Natural, partially observed

* There are several methodological developments to assess causality to some degree (e.g., Granger causality in econometric studies using secondary time series data).

Table 2 – The Clinical Trial Metaphor

Variable	Clinical Trials of Medical Drugs	Randomized Controlled Trials in Marketing
Intervention	Medical treatment (e.g., pill)	Marketing instruments such as message (e.g., TV commercial) on an online media channel
Unit of Analysis	Patient	Consumer
Context	Health care system, current health	Market, consumer expectations
Driver	Genetic make-up	Preference (latent)
Dependent Variable	Health status	Buying attitudes and behavior
Desired outcome	Physiological change	Behavioral change

Table 3 – PMM Outcomes

Type of Finding	Finding in Existing Literature	Reliability
Exploratory Finding	No	Does not hold (or cannot be tested) in most samples
Confirmatory Finding	No	Holds in most samples
Replication	Yes	Holds in all samples

Table 4 – Platforms Used in Each Media Channel and Country

Media Channel	U.S.	China	Netherlands
Display Banner	Major automotive website	-	-
Search Engine	Google	Baidu	-
Social Media	Facebook	Weibo	Facebook
Online Video (Popular TV show)	House	Legend of Zhen Huan	Flikken Maastricht

Table 5– U.S. Sample: Estimates of Response to Advertising

Dependent Measure		Lift in Consideration			
		Funnel Effects		Funnel Effects Per Period	
Constant		0.2607***	(0.04)	0.2640***	(0.04)
Main Effects					
Online video		0.9151***	(0.12)		
Social Media		0.2807*	(0.16)		
Search		0.1005	(0.10)		
Banner		-0.0491	(0.07)		
Interactions					
Online video *x* Social media		-0.0482	(0.25)		
Online video *x* Search		0.0397	(0.17)		
Online *x* Banner		-0.1043	(0.15)		
Search *x* Social media		-0.3110	(0.26)		
Social media *x* Banner		-0.1281	(0.25)		
Banner *x* Search		-0.1015	(0.16)		
Funnel Effects					
4 years until purchase		0.0196	(0.03)	-0.0048	(0.03)
3 years until purchase		-0.0010	(0.03)	-0.0048	(0.03)
2 year until purchase		-0.0550*	(0.03)	-0.0223	(0.04)
Interaction of Time Until Purchase with Treatments - Per Period					
Online video *x*	4 years until purchase			0.2214***	(0.06)
	3 years until purchase			0.3377***	(0.07)
	2 years until purchase			0.3553***	(0.07)
	1 year until purchase			0.9721***	(0.12)
Social *x*	4 years until purchase			0.3351**	(0.15)
	3 years until purchase			0.0738	(0.11)
	2 years until purchase			0.0031	(0.09)
	1 year until purchase			-0.0325	(0.17)
Banner *x*	4 years until purchase			0.0300	(0.05)
	3 years until purchase			-0.0586	(0.06)
	2 years until purchase			-0.1025*	(0.06)
	1 year until purchase			-0.1215	(0.10)
Search *x*	4 years until purchase			-0.0485	(0.06)
	3 years until purchase			-0.0758	(0.06)
	2 years until purchase			0.1202*	(0.07)
	1 year until purchase			0.0729	(0.12)
Controls					
Std. Weekly Usage of Social Media		-0.0161	(0.03)	-0.0144	(0.03)
Std. Weekly Usage of On-Line Video		0.0341	(0.03)	0.0309	(0.03)
R^2		0.0431		0.0470	
Observations		5184		5184	

Dependent variable: lift in consideration of focal brand's car model in segment, measured before and after exposure to treatment. Includes correction for heteroscedasticity when applicable. Time-until-purchase intervals not overlapping (For example, 1 year until purchase includes from 0 to 1 year; 2 years until purchase includes from 1 to 2 years.) Consumers with time-until-purchase higher than 4 years were screened out.

*p<0.10 ** p<0.05 ***p<0.01"

Table 6 – China Sample: Estimates of Response to Advertising

Dependent Measure		**Lift in Consideration**			
		Funnel Effects		**Funnel Effects Per Period**	
Constant		0.0628***	(0.01)	0.0462***	(0.02)
Main Effects and Interactions					
Online video		0.0905***	(0.02)		
Social Media		0.0132	(0.03)		
Search		-0.0057	(0.03)		
Interactions					
Online video *x* Social media		0.0987**	(0.04)		
Online video *x* Search		0.0122	(0.05)		
Search *x* Social media		0.0195	(0.05)		
Funnel Effects					
4 years until purchase		-0.0051	(0.04)	-0.0053	(0.01)
3 years until purchase		-0.0176*	(0.05)	-0.0058	(0.01)
2 year until purchase		-0.0313***	(0.05)	-0.0379***	(0.01)
Interaction of Time Until Purchase with Treatments - Per Period					
Online video *x*	4 years until purchase			0.0284	(0.02)
	3 years until purchase			0.0189	(0.02)
	2 years until purchase			0.0533***	(0.02)
	1 year until purchase			0.1329***	(0.02)
Social *x*	4 years until purchase			0.0402	(0.03)
	3 years until purchase			0.0382	(0.03)
	2 years until purchase			0.0708***	(0.02)
	1 year until purchase			0.0405	(0.03)
Search *x*	4 years until purchase			-0.0297	(0.02)
	3 years until purchase			-0.0094	(0.03)
	2 years until purchase			0.0082	(0.02)
	1 year until purchase			0.0142	(0.03)
Controls					
Std. Weekly Usage of Social Media		0.0010	(0.01)	0.0031	(0.01)
Std. Weekly Usage of On-Line Video		0.0021	(0.01)	0.0022	(0.01)
R^2		0.0741		0.0758	
Observations		921		921	

Dependent variable: lift in consideration of focal brand's car model in segment, measured before and after exposure to treatment. Includes correction for heteroscedasticity when applicable. Time-until-purchase intervals not overlapping (For example, 1 year until purchase includes from 0 to 1 year; 2 years until purchase includes from 1 to 2 years.) Consumers with time-until-purchase higher than 4 years were screened out.

*p<0.10 ** p<0.05 ***p<0.01"

Table 7 – Netherlands Sample: Estimates of Response to Advertising

Dependent Measure		**Lift in Consideration**			
		Funnel Effects		**Funnel Effects Per Period**	
Constant		0.0467	(0.17)	0.2966	(0.22)
Main Effects					
Online video		0.4965**	(0.19)		
Social Media		0.1538	(0.17)		
Interaction					
Online video *x* Social Media		-0.1266	(0.27)		
Funnel Effects					
4 years until purchase		0.0057	(0.05)	-0.1916*	(0.10)
3 years until purchase		0.0191	(0.06)	-0.0451	(0.10)
2 year until purchase		-0.0863	(0.08)	-0.1967	(0.13)
Interaction of Time Until Purchase with Treatments - Per Period					
Online video *x*	4 years until purchase			0.1982**	(0.09)
	3 years until purchase			0.1353	(0.10)
	2 years until purchase			0.2692**	(0.11)
	1 year until purchase			0.2004	(0.24)
Social *x*	4 years until purchase			0.1970**	(0.09)
	3 years until purchase			0.0184	(0.10)
	2 years until purchase			-0.0065	(0.11)
	1 year until purchase			-0.1224	(0.23)
Controls					
Std. Weekly Usage of Social Media		-0.0147	(0.06)	-0.0052	(0.06)
Std. Weekly Usage of On-Line Video		0.0724	(0.06)	0.0632	(0.06)
R^2		0.0186		0.0274	
Observations		796		796	

Dependent variable: lift in consideration of focal brand's car model in segment, measured before and after exposure to treatment. Includes correction for heteroscedasticity when applicable. Time-until-purchase intervals not overlapping (For example, 1 year until purchase includes from 0 to 1 year; 2 years until purchase includes from 1 to 2 years.)Consumers with time-until-purchase higher than 4 years were screened out.

*p<0.10 ** p<0.05 ***p<0.01"

Table 8– Estimates of Response to Advertising, Pooled Data

Dependent Measure		Lift in Consideration			
				Funnel-Specific Effects	
		Channel - Specific Effects	Country - Specific Effects	Simple Pooling	Weighted pooling (unconstrained resid.variance)
Constant		0.2405*** (0.03)	0.2405*** (0.03)	0.3231*** (0.04)	0.4558*** (0.03)
Channel Main Effects					
Online Video		0.9131*** (0.12)			
Social Media		0.2722* (0.16)			
Search		-0.0001 (0.03)			
Banner		-0.0478 (0.07)			
Channel Interactions					
Online video *x* Social Media		-0.0510 (0.25)			
Online video *x* Search		0.0470 (0.17)			
Online video *x* Banner		-0.1035 (0.15)			
Social Media *x* Search		-0.3021 (0.26)			
Social Media *x* Banner		-0.1292 (0.25)			
Search *x* Banner		-0.1023 (0.16)			
Country-Specific Effects					
US					
Online Video US			0.9131*** (0.12)		
Social Media U.S.			0.2722* (0.16)		
Search U.S.			0.0937 (0.10)		
Banner U.S.			-0.0478 (0.07)		
China					
Online Video CN			0.0908*** (0.02)		
Social Media China			0.0143 (0.03)		
SearchChina			-0.0001 (0.03)		
Netherlands					
Online Video NL			0.4993*** (0.19)		
Social Media Netherlands			0.1554 (0.17)		
Funnel-Specific Effect: Interaction of Channel and Time Until Purchase (non-overlapping time intervals)					
Online video *x*	4 years until purchase			0.1927*** (0.04)	0.1070*** (0.03)
	3 years until purchase			0.2690*** (0.04)	0.1038*** (0.02)
	2 years until purchase			0.3103*** (0.05)	0.1264*** (0.02)
	1 year until purchase			0.7166*** (0.07)	0.2323*** (0.02)
Social *x*	4 years until purchase			0.2037*** (0.05)	0.1161*** (0.04)
	3 years until purchase			0.0508 (0.05)	0.0444 (0.03)
	2 years until purchase			0.0231 (0.06)	0.0644*** (0.02)
	1 year until purchase			0.0347 (0.09)	0.0440* (0.03)
Banner *x*	4 years until purchase			0.0577 (0.04)	0.0835* (0.05)
	3 years until purchase			-0.0369 (0.05)	0.0228 (0.05)
	2 years until purchase			-0.0962* (0.05)	-0.0380 (0.05)
	1 year until purchase			-0.0848 (0.08)	0.0146 (0.09)
Search *x*	4 years until purchase			-0.0283 (0.05)	-0.0356 (0.03)
	3 years until purchase			-0.0655 (0.05)	-0.0356 (0.03)
	2 years until purchase			0.1102* (0.06)	0.0451 (0.03)
	1 year until purchase			0.0988 (0.08)	0.0434 (0.03)
Controls					
Std. Weekly Usage of Social Media		-0.0010 (0.01)	-0.0010 (0.01)	-0.0107 (0.02)	-0.0018 (0.01)
Std. Weekly Usage of On-Line Video		0.0146 (0.01)	0.0146 (0.01)	0.0335 (0.02)	0.0181 (0.01)
Fixed Effect China				-0.5984** (0.07)	-0.4696*** (0.04)
Fixed Effect Netherlands				-0.4280** (0.08)	-0.3074*** (0.08)
R^2		0.0678	0.0678	0.0431	0.0476
Observations		6901	6901	6901	6901

Dependent variable: lift in consideration of focal brand's car model in segment, measured before and after exposure to treatment. Includes correction for heteroscedasticity when applicable. Time-until-purchase intervals not overlapping (For example, 1 year until purchase includes from 0 to 1 year; 2 years until purchase includes from 1 to 2 years.)Consumers with time-until-purchase higher than 4 years were screened out.

*p<0.10 ** p<0.05 ***p<0.01"

Table 9–The Exposure-Attention-Engagement Process – Dutch Sample

Dependent Measure:	**Lift in Consideration**					
	Conditional on **Exposure**		Conditional on **Exposure** and **Attention**		Conditional on **Exposure, Attention,** and **Engagement**	
Constant	0.0467	(0.17)	0.0512	(0.17)	0.0960	(0.19)
Main Effects and Interactions						
Online Video	0.4965**	(0.19)	0.7642***	(0.25)	0.7924***	(0.25)
Social	0.1538	(0.17)	0.1923	(0.17)	0.4481*	(0.24)
Interaction						
Online video *x* Social Media	-0.1266	(0.27)	-0.4454	(0.31)	-0.7766**	(0.37)
Funnel Effects						
4 years until purchase	0.0057	(0.05)	0.0030	(0.06)	-0.1113*	(0.06)
3 years until purchase	0.0191	(0.06)	0.0238	(0.07)	0.0210	(0.09)
2 year until purchase	-0.0863	(0.08)	-0.0966	(0.09)	-0.0903	(0.10)
Controls						
Std. Weekly Usage of Social Media	-0.0147	(0.06)	-0.0135	(0.06)	-0.0502	(0.07)
Std. Weekly Usage of On-Line Video	0.0724	(0.06)	0.0757	(0.07)	0.0763	(0.07)
Observations	796		714		509	
R^2	0.0186		0.0262		0.0348	

*p<0.10 **p<0.05 ***p<0.01. Includes correction for heteroscedasticity when applicable.

Table 10 – Key Substantive Findings

Type of Finding	Finding	Finding in Existing Literature	Hold in All Tested Samples
Exploratory Finding	Engagement and attention strengthen consumer response to social media and online video	No	Yes, but only testable in one sample (extra control variable)
Exploratory Finding	Negative interaction of response to video and to social media under high levels of engagement	No	Yes, but only testable in one sample (extra control variable)
Confirmatory Finding	Response to social media is stronger at earlier stages of consideration	No	Holds in most samples
Replication	Strong response to online video advertising holds in YouTube videos	Yes	Yes

FIGURES

Figure 1 –Measuring Media Effectiveness: The *Exposure-Attention-Engagement* Process

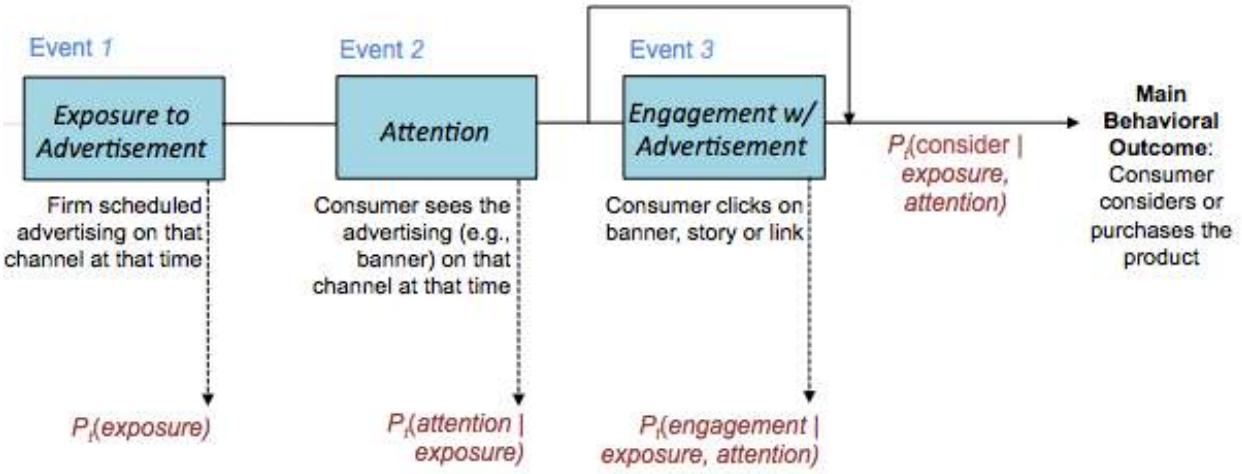
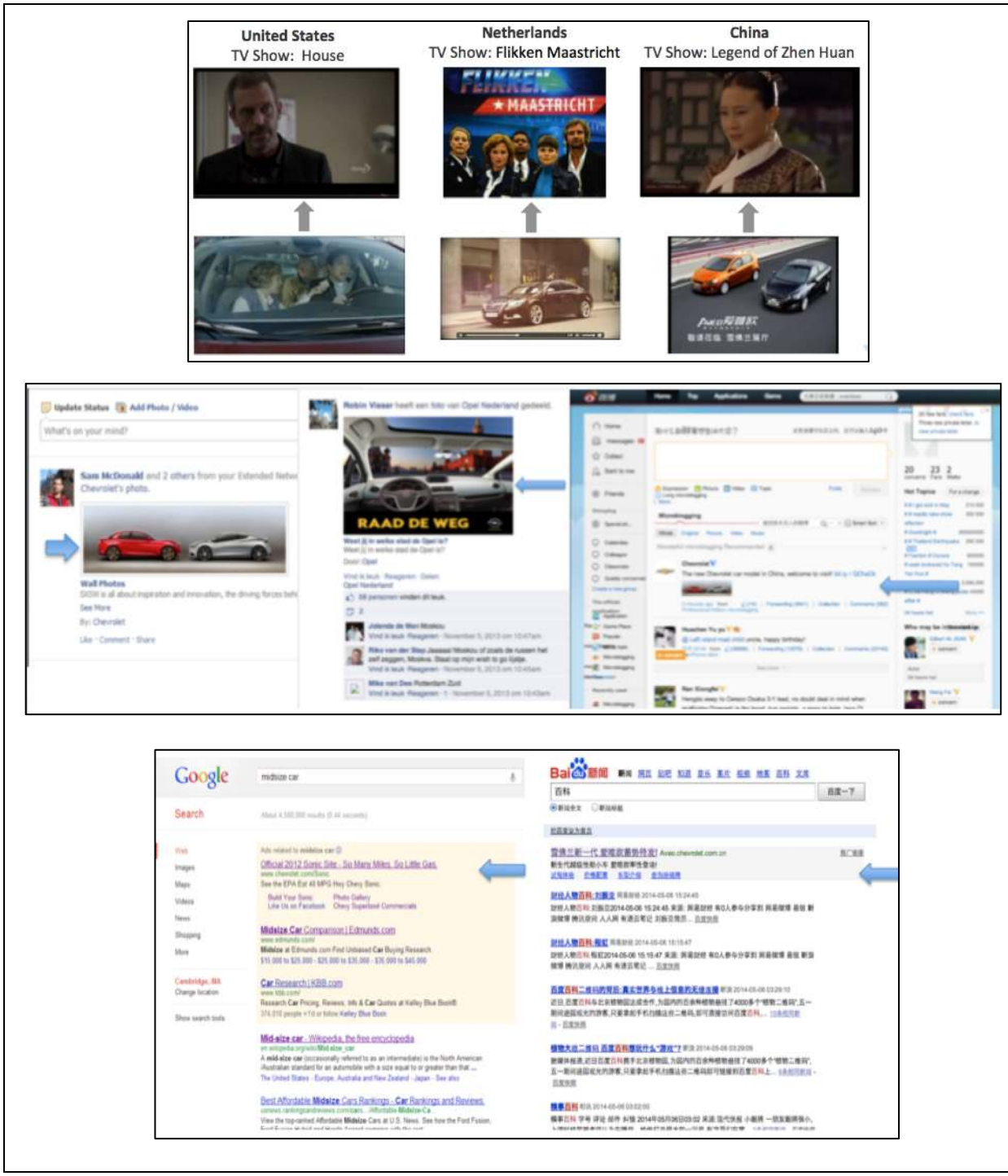


Figure 2 – Treatments: TV Shows (top), Social Media (middle) and Search (bottom)

.



APPENDIX A – Checks on Potential Threats to Validity

A key source of self-selection that potentially could affect our study is that respondents that decided to not install our Chrome plug-in, i.e., the non-compliers, could have different behavior from the set of respondents that decided to comply. The concern is that those who drop out are different from those who stay. While we did use randomized allocation to the various treatments and control in every country, self-selection could promote systematic differences in how dropouts respond to online media advertising because those who stay could perceive technology differently in such a way that they are differently affected by online advertising. In the first study we re-routed non-compliers to cells that did not require chrome and checked for statistically significant differences on consideration between non-compliers and respondents that were originally randomly assigned to the cells that did not require chrome. We did not find statistically significant differences. This suggests non-compliers were not affected by online advertising differently from compliers. In the third study we further inspected selection using a different approach by collecting additional information from non-compliers to explicitly model the decision to download Google Chrome, and correct our estimates accordingly.

We also measured the amount of hours each consumer spent per week on average on mobile telephone, tablet, and computer (laptops or desktops). These three questions about usage of technology were asked before the respondents were invited to download and install our Chrome plug-in. Let $Std.WeeklyUsageofIT_i$ be the mean-centered standardized average number of hours per week consumer *i* uses these three technologies. In our selection model we assume the decision to not download the Google Chrome extension is likely to be influenced by consumer's expertise with information technology. We also assume that these skills are unlikely to substantively change consumer response to online video advertising and social media advertising. Not being able to install a browser plugin or being a seasoned manager of computer networks can affect how likely a person would install a browser extension but it is unlikely to have an impact on how advertising affects her consideration behavior. Hence we have a selection model with at least one variable that does not belong to the main model, as follows:

$$Compliance_i = \propto + \beta_1 Std.WeeklyUsageofITi_{\ i} + \beta_2 Std.Gender_i + \varepsilon_i \qquad \text{(A1)}$$

We applied Equation 2 to the set of respondents that were randomly assigned to the social media cell, which requires the download and installation of our plug-in. We then inspected how well our selection model predicts the decision to accept or reject the invitation to download and install the Chrome plug-in. Our selection model performs extremely well – it can correctly predict 88% of the binary decisions of whether to comply.

Having established its validity, we used the selection model to Heckman-correct our estimates in the Dutch sample. We expected the corrected estimates would not substantially differ from our non-corrected estimates because we do not anticipate non-compliance to be a major issue given the widespread usage of Google Chrome browser. The first column of Table 10 shows our baseline model without correction. The second column shows the Heckman-corrected estimates using Equations 1 and 2.

Table A1 – Analysis of Self-Selection Due to Chrome Plug-In Download Decision in the Dutch Study

Dependent Measure:	Lift in Consideration			
	Baseline		With Selection Model	
Constant	-0.0303	(0.12)	0.2749	(0.19)
Treatments				
TV	0.5313***	(0.19)	0.5315***	(0.19)
Social	0.2031	(0.17)	0.2108	(0.17)
Social & TV	-0.1649	(0.26)	-0.1638	(0.26)
Controls				
Std. Weekly Usage of Social Media	-0.0096	(0.05)	-0.1638	(0.26)
Std. Weekly Usage of On-Line Video	0.0561	(0.05)	-0.0159	(0.05)
Selection Model				
Constant			0.1177	(0.09)
Weekly Usage IT			0.1654***	(0.03)
Gender			-0.3252***	(0.06)
Arc Tang(Rho)			-0.1722	(0.09)
Ln(Sigma)			0.6173***	(0.07)
Observations	796		2079	
Censored Observations	-		1,283	
Uncensored Observations	796		796	

* p<0.10 ** p<0.05 ***p<0.01". Includes correction for heteroscedasticity when applicable. Correction for self-selection on binary decision to download Chrome based on Heckman (1979). Wald test of indep. eqns. (rho = 0): chi2(1) = 0.02 Prob > chi2 = 0.8991

As expected, the estimates do not change substantially and we fail to reject the hypothesis that $\rho=0$, suggesting these OLS estimates are not biased. A second self-selection concern could be raised regarding the decision to login into the social media platforms, visit the website or do a search upon receiving the invitation. However, this problem was preempted with clear instruc-

tions, incentive alignment and trimming at the analysis level i.e., we conditioned all analysis and estimates on engagement with channel (see event 1 in Figure 1). We removed respondents that did not minimally use the channel and inspected the effectiveness of the advertising on the channel conditional on consumers using the channel.

A third potential threat to validity could come from unobserved sources of heterogeneity. Some across-study unobserved heterogeneity is captured in the sample-specific country fixed effects, used on Table 10. A fourth potential threat to validity comes from unobserved sources of other dependencies across observations, for example stemming from some respondents being more susceptible to media censorship.

APPENDIX B – Screening, Weighting, and Estimation Procedures

Respondents were screened for the category of automobiles they were interested in. In the U.S. we had stimuli for the following categories: small, compact, and mid/full-size. In Europe we focused on small, medium, and large. In China we focused on micro, small, medium, and mid/full-size. Respondents were also screened for minimal involvement with the media channel (measured as the number of clicks on media channel, e.g., at least one click on the social media platform), time spent on the treatment (at least 90% of the duration of the assigned TV show), and the time window in which they planned to buy a car (4 years or less). Respondents were also screened for age (e.g., at least 18 years old), time taken to complete the questionnaire (10% trimming), and to make sure they did not work for companies in industries directly related with the study (the excluded industries are advertising, automotive dealerships, manufacturer or suppliers to manufacturers, press, radio, TV or journalism, and market research).

Pooling and estimation.

There is clearly much value in pooling together data across studies but direct pooling is often not straightforward due to the constraint on residual variances of the samples. This constraint can be corrected by weighting the observations proportionally to the inverse of the variance of the group each observation belongs to (Greene, 2011). This correction also accounts for differences in sample sizes because larger samples yield lower variance. Such a correction can be easily done with the standard analytical weight option on modern statistical packages. We adapt Equation 1 to estimate main effects, interactions, and funnel effects at the country level, and we apply this weighting strategy to pool data across the three studies with potentially unequal variances and sizes. In total, the pooled dataset amounts to 6,901 observations. Pooled data also allow us to directly compare estimates across countries. Pairwise Chow tests on country-specific

estimates using the weighted dataset shows the online video estimates are the same for China and US (F = 16.96, p = 0.000), China and the Netherlands (F = 4.85, p = 0.028), and US and the Netherlands (F = 7.35, p = 0.01).

We estimated the models in U.S., China, and Netherlands in Tables 5, 6, 7, 8 (columns 1,2 and 3), and 9 with the Stata command 'regress' and the option 'vce(robust)' to control for heteroskedacity when applicable We estimated the weighted model in the last column of Table 8 in three steps. First, we run the variance-constrained regression. Second, we generate the weights based on the residuals of each group. Third, we run the variance-unconstrained regression with these weights. The Stata pseudocode is as follows[1].

*** Step 1**

regress lift <*all independent variables*>

*** Step 2**

predict r_ch, resid

gen w_ch = 0

sum r_ch if fixed_effect_nl = =1

replace w_ch = r(var) if fixed_effect_nl = =1

sum r_ch if fixed_effect_cn = =1

replace w_ch = r(var) if fixed_effect_cn = =1

sum r_ch if fixed_effect_us = =1

replace w_ch = r(var) if fixed_effect_us = =1

*** Step 3**

regress lift <*all independent variables*> [aw=1/w_ch], vce robust

[1] Further details are also available at http://www.stata.com/support/faqs/statistics/pooling-data-and-chow-tests/

Appendix C – Early Stopping and Interim Analysis

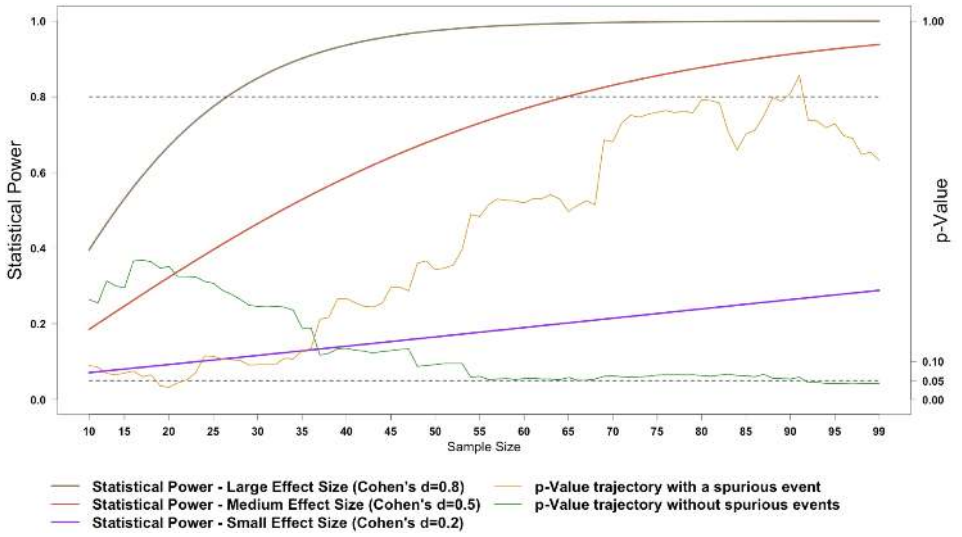
This appendix has two parts. First, we provide background information and a conceptual discussion on early stopping and interim analysis. Second, we show how we planned, documented, and implemented early stopping in our study.

Considerations on Early Stopping and Interim Analysis

Though adaptive allocation of sample to treatments in general (and in particular, early stopping) is a key contributor to the power of PMM, it is important to note that we do not advocate stopping data collection for a treatment cell at the first signs of statistical significance. Stopping at first signs of significance would reduce replicability, partly because of spurious significance (Simonsohn and Simmons, 2014). Instead, we strongly encourage every application of Prospective Meta-Analysis to register all decisions regarding sampling in a protocol in advance, before data are collected. A possible criterion is to stop a cell once enough power has been achieved for the expected effect size given past studies. The indications needed for sampling decisions can be based on interim analysis of p-curves, effect sizes and statistical power. Therefore we apply the same rigor to early stopping in PMM as clinical researchers do in medical sciences. As in medicine, all sampling decisions need to be discussed, planned and explicitly documented before the trial starts[c].

Figure C1 shows typical interim p-value curves and statistical power for different effect sizes at each data point of a synthetic dataset. Note that the curves in Figure C1 are sketched based on synthetic data for the sole purpose of illustrating the concept of spurious significance.

Figure C1: Interim Analysis Helps Avoid Spurious Significance.



Interim analysis of p-curves allows the researcher to visually inspect the consistency between statistical power and p-values trajectories, and check the stability of p-values over sample size. Inconsistencies suggest spurious statistical significance due to lack of statistical power. For example, the yellow p-value curve in Figure C1 shows that an estimate that seems to be statistically significant at n=20 becomes non-significant soon after that. Hence, having stopped early would have been a mistake. On the other hand, the green p-curve shows a different pattern – after reaching the critical threshold of p=0.05, it remains below that level as more sample is added, suggesting no spurious significance. The plot also shows that sufficient power to detect medium effect sizes based on the conventional 0.8 power threshold is only achieved at n=60, which can be used by the data analyst as a sanity check, searching for signs of underpowered significance. These simple statistical results and checks can help researchers prevent spurious significance, and can be easily computed. The above statistical power curves were computed for each sample point and each of Cohen's effect sizes (d=0.2, d=0.5 and d=0.8) using the conventional significance level (p=0.05), and two-sided, two-paired-samples t-tests implemented by the pwr.t.test()

R package. For more detailed discussions on the principles underlying p-value curves please refer to Simmons and Simonsohn (2011) and Simonsohn and Simmons (2014).

Early Stopping of the Online Video Cell in Study #3

We decided to implement early stopping in our third study because the results of the first two studies (in the U.S., and China) provided us with sufficient findings for strong expectations regarding the effects of online video ads in Netherlands. We discuss our application of early stopping in two steps. First, we show how we planned and documented the decision to stop assigning respondents to the online video experimental cell in study 3. Second, we show how we monitored the data collection of that study until that cell stopped receiving respondents.

Planning and Documenting. We decided that we would stop assigning respondents to the online video experimental cell after the results from the U.S. were replicated. We needed some approximate idea of what would be a reasonable (yet conservative) expectation of when this could happen. Thus, we estimated the sample needed to replicate such results with a simple statistical power analysis calculation (for convenience, we used the G*Power software available from www.gpower.hhu.de/en.html).

Figure C2 – Sample Requirement to Replicate U.S. Online Video Effect in the Dutch Sample

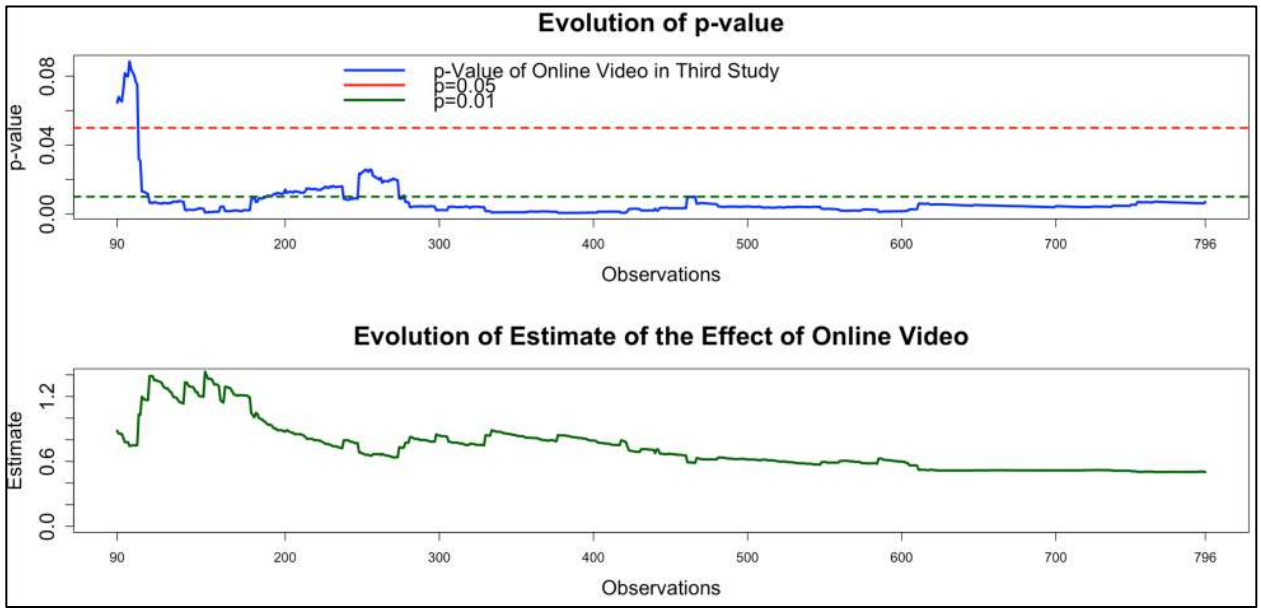
```
Analysis:   A priori: Compute required sample size
Input:            Effect size f²                    =     0.3
                  α err prob                        =     0.05
                  Power (1-β err prob)              =     0.95
                  Number of predictors              =     14
Output:           Noncentrality parameter λ         =     31.2000000
                  Critical F                        =     1.8044774
                  Numerator df                      =     14
                  Denominator df                    =     89
                  Total sample size                 =     104
                  Actual power                      =     0.9521440
```

Figure C2 shows that we could expect that the strong online video results found in study 1 would be found after the online video cell in study 3 had achieved about 100 subjects. We also

decided that instead of immediately stopping assigning respondents to the online video experimental cell at the 100th subject, we would monitor the evolution of p-values before and after we had achieved sufficient statistical power to replicate U.S. results, as described next.

Monitoring and Stopping. We decided we would monitor the estimates and p-values as more and more respondents completed the experiment. Figure C3 shows the evolution of the p-curve after each of the 796 respondents completed the questionnaire.

Figure C3: Evolution of p-Values and the Estimate of the Effect of Online Video Ads on Lift in the Dutch Sample



The first plot in Figure C3 shows that once the p-value crossed the 0.05 threshold, it stayed below that threshold throughout the remaining of the study. The second plot shows that the estimate stabilized after the 600th respondent of that study, in total.

Together, both plots and the power analysis calculations allowed us to minimize sample costs and increase sample in cells with treatments that have weaker effects.