

KEYWORDS

micro-targeting  
disclosure labels  
trustworthiness  
credibility  
persuasion knowledge

RESEARCH PAPER

## Empty Transparency?

### The Effects on Credibility and Trustworthiness of Targeting Disclosure Labels for Micro-Targeted Political Advertisements

Martin-Pieter Jansen<sup>1\*</sup>  \ Nicole C. Krämer<sup>1,2</sup> 

<sup>1</sup> Social Psychology: Media and Communication, University of Duisburg-Essen

<sup>2</sup> Research Center Trustworthy Data Science and Security

\*Corresponding author, [martin-pieter.jansen@uni-due.de](mailto:martin-pieter.jansen@uni-due.de)

ABSTRACT

Political micro-targeting describes the use of data to identify members of a target audience and send messages designed to fit their views and resonate with them. The practice has received considerable attention of late, especially around questions of transparency. This study explores one potential solution to this quandary, namely, disclosure labels. Adopting a pre-registered online one-factorial three-group between-subjects experimental design, we have investigated how different types of disclosure labels for micro-targeted advertisements impact source and message credibility, as well as source trustworthiness. Furthermore, we have investigated the potential mediating effect of persuasion knowledge on these effects. We exposed 227 German Facebook users to either a Facebook advertisement without a disclosure label, a sponsored disclosure label, or a targeting disclosure label that stated they were targeted based on their online behavior. The results demonstrate small and non-significant differences between groups regarding source and message credibility and source trustworthiness, with no mediation by persuasion knowledge observed. Additionally, most participants did not recall the disclosure we exposed them

to, potentially explaining these small effects within our sample. In conclusion, our targeting disclosure approaches were insufficiently informative. Hence, we argue that platforms should put more effort into improving transparency for their users than they currently do.

## 1 Introduction

In recent years, political campaigns have moved with their audiences from face-to-face contact to a focus on digital connections and interactions and relationship development. In the current digital media landscape, among the most common ways that political parties contact potential voters is through social networking sites (SNSs) (Giasson et al., 2019). Parties and political consultancy firms use these platforms to direct their messages to the target groups that they predict to be most susceptible to them. The information parties use to target these individuals is oftentimes a product of their online behavior (Matz et al., 2017). The large volumes of information that platforms gather from usage and the information that users voluntarily share – such as demographics, likes, interests, and location data – enable these targeting activities to be personalized and, in some cases, extremely narrowly targeted (i.e., at small groups of people with the same interests) (Dobber et al., 2017). These messages are developed to resonate with the specific target audience. The demarcation and targeting based on this data describe the concept of *micro-targeting* or, in the case of politics, *political micro-targeting* (PMT) (Kruikemeier et al., 2016; Zarouali et al., 2020; Zuiderveen Borgesius et al., 2018).

In the most generally known illustration of PMT, political consultancy firm Cambridge Analytica allegedly gathered and used data from more than 50 million Facebook users to establish psychological profiles and target users with messages that would persuade them as strongly as possible (Cadwalladr, 2018). According to whistleblowers, this contributed to Donald Trump's presidential victory and the Leave campaign's success in the Brexit referendum (Kaiser, 2019; Wylie, 2019).

Although PMT can mobilize potential voters that might otherwise have been left out, personalized content and the sharing of relevant information (Zuiderveen Borgesius et al., 2018) imply risks for society and democracy. PMT allows senders of messages to contribute to selective information exposure, which sees parties present themselves as single-issue parties to different individuals (with different issues). In such cases, PMT could lead to biased perceptions regarding parties if voters are not aware of other issues that the party focuses on, potentially threatening the marketplace of ideas within our democratic society (Barocas, 2012; Zuiderveen Borgesius et al., 2018). Furthermore, PMT could deliberately ignore certain target groups: Because it is possible to

mobilize voters that are more likely to vote for a certain party, those who would likely vote for the other parties are purposely neglected, decreasing scrutiny of the democratic process (Jamieson, 2013). In turn, this could expand the gap in representation in governments by making those who are targeted more strongly represented (Endres & Kelly, 2018). Finally, PMT could undermine the public sphere by helping to shield potential voters from information or viewpoints that might challenge their beliefs and values (Gorton, 2016, p. 69).

One solution to these risks could be increasing transparency by using disclosure labels. These labels are often used for regular marketing activities – such as advertorials (Boerman & van Reijmersdal, 2016), sponsored vlogs (Van Der Goot et al., 2021), and influencer marketing (van Reijmersdal et al., 2020) – and could represent a chance to inform the public of the nature of advertisements and the targeting that took place. Simultaneously, disclosures could provide users with more information about the party that pays for the advertisement, which could differ from the political party itself. In an example of a new regulatory approach, the EU Digital Services Act states that targeted information and advertisements should include information about when and on whose behalf content is displayed (European Commission, 2022).

In one of the first experiments on the use of disclosure labels on PMT advertisements, Kruikemeier et al. (2016) showed that if individuals notice disclosure labels on a micro-targeted Facebook post, they will better understand that the post is an advertisement. This promotes the activation of *persuasion knowledge* (PK), a mechanism that helps users identify persuasion attempts based on previous exposure and experience and can be influenced by awareness of the advertisement's topic and knowledge of the message's sender (Friestad & Wright, 1994; Jung & Heo, 2019). Furthermore, PK has been shown to effectuate a more critical style of processing a message, which might be problematic for the sender of a message, especially when it is a political party (Boerman & Kruikemeier, 2016; Campbell, 1995; Main et al., 2007; Wentzel et al., 2010).

Building on the work of Kruikemeier et al. (2016), this work investigates the effects of different types of disclosures on source trustworthiness while also investigating the potential mediating role of PK on these effects. However, we move beyond the measures used by Kruikemeier et al. (2016) by also including source and message credibility as dependent variables. Both of these constructs are found to be important predictors in the voting process (Carr & Hayes, 2014; Hetherington, 1999; Housholder & LaMarre, 2014; Madsen, 2019; Main et al., 2007). Furthermore, we simulate PMT differently by letting participants indicate whether they agree with certain statements before exposing them to our conditions, thereby aligning advertisements with their beliefs. In addition, we use a different disclosure label for our targeting disclosure. This label is in line with the disclosure labels Facebook currently uses but includes more salience about the targeting practices being used. Hence, we arrive at our study's central research question:

How does placing targeting disclosures labels above micro-targeted political advertisements impact source credibility, message credibility, and source trustworthiness, and what is the mediating role of persuasion knowledge on these effects?

## 2 Theoretical Background

### 2.1 Micro-Targeting

SNSs such as Facebook and Instagram and search engines such as Google or Bing offer considerable marketing potential. Constant usage and online behavior enable users of these sites to help create databases that are perfect for brand promotion (Barbu, 2014). For senders of messages, these databases provide detailed information about what a user likes and dislikes, which messages they are more given to interacting with, and, hence, which messages are potentially more influential (Winter et al., 2021; Yan et al., 2009). Companies and other parties that utilize advertisements can target specific messages at specific target groups where those messages are likely to be more effective, a practice known as *behavioral targeting* (Matz et al., 2017; Yan et al., 2009). At a time when people are more online than ever before, leaving breadcrumbs as they go from website to website, this practice has become more efficient than ever. These breadcrumbs that people unconsciously leave behind include information about their personal lives and their (online) behavior and can sometimes even be used to predict personality traits based on previous interactions with websites and other forms of online content (e.g., Facebook likes) (Kosinski et al., 2013; Matz et al., 2017; Yan et al., 2009; Zarouali et al., 2020). Furthermore, by applying intensive algorithms, senders can automatically cluster users that share attributes by using machine learning to process these new types of data (Papakyriakopoulos et al., 2018). According to Wilson (2017), this is something that could be automated further by having artificial intelligence processes move from personality profiling to specialized content generation and delivery.

Within the political realm, the goal is not to sell products but to sell the story and ideology of a politician or a party to receive votes. With techniques borrowed from commercial companies and marketing agencies, political actors aim to persuade voters that are unsure about their partisanship, voters that are unsatisfied with previously preferred candidates, and potential new voters. Of course, the concept of dividing different types of potential voters and targeting them with certain messages is not new: Before the internet, parties used canvassing strategies that relied on different fliers in different states or for different zip codes (Barbu, 2014; Gandy, 2000; Murray & Scime, 2010). However, the amount of data gathered by platforms and advertisers makes it possible

to target smaller groups, build look-a-like audiences, and even psychological profiles based on groupings such as the Big Five personality traits (Zarouali et al., 2020; Zuiderveen Borgesius et al., 2018). Although the goal may differ, the means of achieving that goal very much resembles the approach used by advertising agencies (Dobber et al., 2017). The use of very precise targeting data and tailored political messages, developed to resonate more effectively within specific target groups on SNSs and search engines, is often referred to as PMT (Barbu, 2014; Endres & Kelly, 2018; Zuiderveen Borgesius et al., 2018).

Communication strategies like PMT remain interesting for political parties because of the low costs and the potential to engage with as many potential voters as possible. To initially implement these strategies, few resources are necessary, which combines with scalability to make the approach all the more attractive to political parties (Dobber et al., 2017). PMT allows senders of messages to reach any individual or group with any message. The party that pays the most money for the placement of an advertisement “buys” the attention of the audience without regard for the potential harm to democracy (Bodó et al., 2017). PMT enables the campaign employees that previously retrieved knowledge about targets from focus groups to know exactly what buttons to press to obtain the desired result, namely, in this case, votes (Gorton, 2016). Campaign employees can target the groups more likely to vote in favor of their candidate. In turn, targeting only citizens from groups with a high probability of voting can increase the gap between citizens and their representation in government (Endres & Kelly, 2018).

Since Trump’s run for office and Brexit, little has changed. Neither journalists nor politicians nor researchers know exactly the extent of Cambridge Analytica’s influence, meaning not only the ethical aspects of the company’s work but also its potential societal impact remain subject to debate (Ortega, 2022; Zuiderveen Borgesius et al., 2018). For example, it is hard for researchers to replicate or simulate micro-targeting because the practice is a *black box*: It is impossible to know precisely what is happening at political consultancy firms or how their algorithms and models operate.

However, there is work that investigates the effects of targeting at different levels. In a randomized field experiment at the zip-code level of targeting, Coppock et al. (2022) found that digital political advertising demonstrated very small estimated effects on vote share. In a study on PMT specifically, Liberini et al. (2020) produced different results regarding the 2016 US elections. For instance, they demonstrated that neutral voters who accessed Facebook daily were up to twice as likely to support the Republican candidates than neutral voters who did not have a Facebook account. This aligns with the idea that voters who are less sure about their partisanship are the most interesting for campaigns because they are less likely to keep voting for the same party, making them more likely to vote for a different campaign’s candidate (Endres & Kelly, 2018).

## 2.2 Countermeasures

However, instead of focusing on these roadblocks and the sender side of PMT, the SNSs that consultancy firms use to influence their audiences could represent a solution. Researchers use different countermeasures to minimize the potentially harmful effects of PMT described earlier in this work. One starting point for transparency could be the development of *ad archives or ad libraries* that store all of the advertisements that have run on an SNS (including potential targeting measures) stored and make them openly available to the public (Leerssen et al., 2019). This is something that Meta provides with its ad library and CrowdTangle platform.<sup>1</sup> Nonetheless, Meta has received criticism about these tools not being accessible to everybody and the fact that the company itself manages the mechanisms (Elswah & Howard, 2020). Ben-David (2020) has even argued that Meta's sustained control over this public data enables the company to keep citizens away from information until its contents become the past. Other commentators have suggested that until the companies behind SNSs start regulating the political messages that appear on their platforms, the best solution for voters is to rely on themselves, be cautious, and check their information diet, which the aforementioned tools might make possible (Ghosh, 2018).

Among the most transparent approaches to countering these potentially harmful effects – and an approach that seems convenient – involves implementing targeting disclosure labels (Binford et al., 2021; Kruikemeier et al., 2016). Targeting disclosure labels could be a perfect middle-ground between total platform transparency and user self-reliance. Platforms need not give up their precise targeting information – a substantial part of their business models – and users need not invest in becoming so self-reliant. By providing users with information regarding the nature of an advertisement, platforms demonstrate the transparency that seems critical for users to distinguish between different types of messages and advertisements (Amazeen & Wojdyski, 2020; Binford et al., 2021).

Furthermore, when advertisers undertake covert information-collection strategies for targeting purposes, consumers may experience feelings of vulnerability. Informational cues could be used to offset these effects (Aguirre et al., 2015). In recent attempts to be more transparent about the nature of advertisements that look like general content in a newsfeed, some SNSs and search engines have started using disclosure labels on advertisements and other sponsored content to show that a sender paid to place the message in question (Binford et al., 2021; Jung & Heo, 2019; van Reijmersdal et al., 2016). Disclosure labels on advertisements exist in many forms. Within the realm of marketing, more prominent disclosure labels are seemingly more effective. Prominent disclosure labels are those that are more easily seen by users and therefore lead

---

<sup>1</sup> CrowdTangle is a public insights tool from Facebook that gives people with access (i.e., journalists, researchers, and social media professionals) insights into public content on the platform (Bleakley, n.d.).

to higher levels of recognition that an advertisement is indeed an advertisement (Amazeen & Wojdyski, 2020). The level of prominence has been found to be important because users are less likely to recognize advertisements when impartial disclosure labels are used (i.e., labels that vaguely state something is sponsored, or just use an icon without explanations; Stubb & Colliander, 2019).

Existing research has not observed differences in the length (in seconds) of exposure to a disclosure label (Boerman et al., 2012). The source of a persuasive message is perceived to be more credible when explicit disclosure labels are used, that is, labels with complete and exact descriptions of what is sponsored and by whom. In such cases, users are aware of potential biases of the source and can integrate that information into their perception of the source and counter uncertainty regarding the source's intentions (Carr & Hayes, 2014). Nonetheless, recent research on disclosure labels for advertisements on SNSs shows that disclosure labels are likely to negatively impact the credibility of both the source and the message (Deng et al., 2020). However, these works both investigated disclosures in the context of regular advertising and sponsored content, which does not incorporate the same levels of personalization or utilize user data to show them specific advertisements. This is pertinent because other work has demonstrated that personalization can be perceived as intrusive (Segijn & van Ooijen, 2022) or even creepy (De Keyzer et al., 2022).

Perceived credibility influences how we use and process information from a particular source (Madsen, 2019). Credibility and trust are more important when the subject of discussion is something the receiver of a message has less knowledge about and when they lack access to information regarding that subject. Although politicians are oftentimes considered experts who specialize in a certain subject, their knowledge will be understood as more valid if they are perceived as more credible and trustworthy. Furthermore, trust directly influences the choice of a political candidate (Hetherington, 1999) and increases the intention to vote for them (Housholder & LaMarre, 2014). This means that when a potential voter deems a candidate more trustworthy and capable, they are more likely to vote for them. Therefore, credibility and trustworthiness both represent important constructs, especially in political communication. Furthermore, when communicating via persuasive messages on SNSs, the receiver's recognition of the persuasion attempt is also important. Understanding the persuasive intent of a message might undermine the trustworthiness of the sender because their aims become known (Main et al., 2007). Furthermore, disclosure labels regarding a message's sponsorship influence perceptions of the opinion leader's (e.g., a political actor) credibility (Carr & Hayes, 2014).

Instead of regular disclosure labels, this study focuses on targeting disclosure labels. These disclosure labels provide users with more information about the fact that they have been targeted by a sender of a persuasive message. This is done while remaining consistent with how SNSs design their regular "sponsored" disclosure labels. This study distinguishes between Facebook posts

without disclosure labels, Facebook advertisements or posts, with the regular disclosure labels that the platform currently uses, and posts featuring a targeting disclosure label that provides users with more salient information about the targeting practices that have taken place. Furthermore, because the targeting disclosure label includes more words, users might notice it more. This could increase the central processing of the disclosure label, encouraging better comprehension of the targeting occurring (Kruikemeier et al., 2016; Petty & Cacioppo, 1986). Although the extant studies previously discussed have demonstrated both increases and decreases in credibility via the use of disclosures, we expect that the transparency that disclosure provides about the targeting and personalization taking place (i.e., the use of user data) decreases the credibility of the source. Therefore, we focus on targeting disclosure labels and investigating their effects on the credibility of the source, the credibility of the message, and the trustworthiness of the source. Given multiple definitions of source and message credibility and source trustworthiness exist, we would like to emphasize the interpretations used in this work. We understand source credibility according to the work of Winter and Krämer (2014), who use competence as an operationalization of credibility that “indicates if the sender is *able* to provide valid statements on a topic” (p. 437). Next, we use the work of Appelman and Sundar (2016) to understand message credibility as “an individual’s judgment of the veracity of the content of communication” (p. 5). Finally, we interpret source trustworthiness as the honesty of the communicator, following its use in existing research around PMT (Kruikemeier et al., 2016). Taking the above into account, we assume that:

H1: A micro-targeted political Facebook message with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will lower (a) the perceived credibility of both the source and the message and (b) the perceived trustworthiness of the source.

## 2.3 Persuasion Knowledge

Modern internet users are exposed to many advertisements and influential messages. Previous exposure allows them to develop personal beliefs and knowledge about those messages in order to manage them. In the context of classic persuasion attempts, this has encouraged the development of the term PK, defined as personal knowledge and beliefs about the motives and tactics of an advertisement and the advertisement’s source (Friestad & Wright, 1994). The development of PK depends on previous exposure to and experiences with advertisements and persuasive messages. Additionally, knowledge of the topic and knowledge of the message’s sender (also known as the agent) both influence PK (Friestad & Wright, 1994; Jung & Heo, 2019). In one of the first studies on targeting disclosure labels for PMT messages on Facebook, Kruikemeier et al. (2016) found that if users notice the disclosure labels on a PMT Facebook

post, they better understand that the Facebook post is an advertisement sent by a political party, activating PK. Targeting disclosure labels can be considered an attempt to inform users about the targeting being practiced. Furthermore, a disclosure label will activate PK, which advertisements without disclosure labels are less likely to do (Boerman et al., 2012; Kruikemeier et al., 2016). Based on these findings, we propose our second hypothesis:

H2: A micro-targeted political Facebook advertisement with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will activate stronger persuasion knowledge.

PK has also been shown to lead users to a more critical style of processing. This, in turn, influences the sender's evaluations, possibly lessening their perception of sincerity and trustworthiness (Boerman & Kruikemeier, 2016; Campbell, 1995; Main et al., 2007; Wentzel et al., 2010). Targeting transparency is not substantially addressed by existing regulations: There is no limit to the extent of targeting and only limited oversight concerning the information used to target users (Dommett, 2020). Beyond this lack of regulation, a gap in the literature regarding transparency disclosure labels for PMT is apparent. Some authors argue that previous research regarding PMT suffers from mono-theoretical blindness, providing only an overview of regulations or focusing on campaign practices or big data analytics (Bodó et al., 2017; Zuiderveen Borgesius et al., 2018). This has prompted calls for more studies integrating more theoretical concepts and providing more insights into PMT-related questions (Bodó et al., 2017).

In its current form, PMT is a recent development in the context of political advertising and campaigning, and there is limited research concerning solutions that can counter PMT and inform users about micro-targeted messages. Nonetheless, some researchers have acknowledged concerns about the potential harms to democracy and the marketplace of ideas that PMT could ultimately represent (Barocas, 2012; Gorton, 2016; Zuiderveen Borgesius et al., 2018). For example, the already discussed work of Kruikemeier et al. (2016) revealed that respondents appeared to resist personalized content when they noticed a disclosure label. Furthermore, they concluded that the opportunity of personalizing ads to reach possible voters might not always be as beneficial in practice, leading them to call for future work investigating negative implications, such as political content avoidance. Based on these findings, we attempt to further investigate potential differences in message credibility, source credibility, and source trustworthiness associated with the mediating role of PK. Although research on disclosure labels, PK, source trustworthiness, and credibility has been conducted in the context of marketing activities (Carr & Hayes, 2014; Deng et al., 2020; Stubb & Colliander, 2019), it is important to investigate these presumed relationships within the context of PMT. Additionally, because both PK activation and transparency disclosure labels have been demonstrated to decrease credibility and trustworthiness in the regular marketing context, it

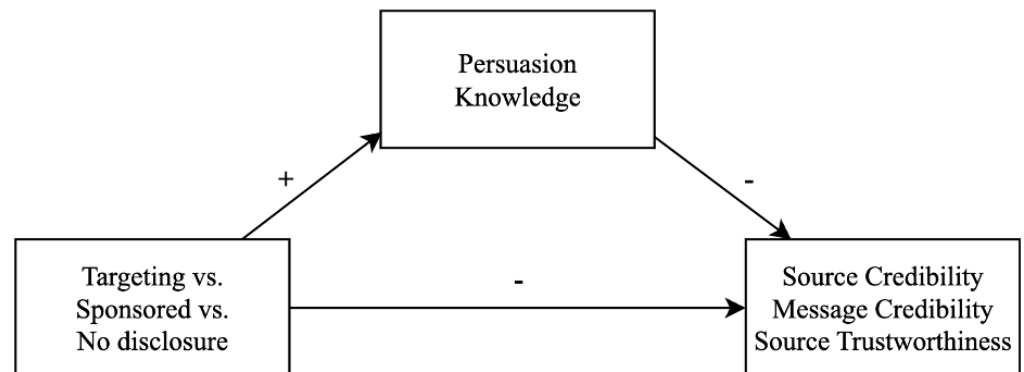
is critical to interrogate the possible existence of these undesirable effects in the political context. Furthermore, research has demonstrated that lower credibility and trustworthiness tend to decrease willingness to vote for a certain party or candidate, making such practices counter-productive for political actors (Hetherington, 1999; Housholder & LaMarre, 2014; Madsen, 2019). Based on these findings, we propose our third hypothesis:

H3: A micro-targeted political Facebook advertisement with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will activate stronger persuasion knowledge, negatively impacting (a) source and message credibility and (b) source trustworthiness.

Returning to the work of Kruikemeier et al. (2016), that study saw them examine the relationship between exposure to personalized political ads on Facebook and voter intention to engage in electronic word of mouth and the perceived trustworthiness of those ads. They also investigated the mediating role of PK within this relationship. During their experiment, they exposed respondents ( $N=122$ ) to either a regular Facebook message, a Facebook advertisement with a disclosure label, or a Facebook message with a disclosure label and an explanation about personalized advertising on Facebook. Within their sample, they found that exposure to a personalized ad from a political party activates PK. Although this, in turn, lowers the intention to engage in electronic word of mouth, this only holds for participants that recall the disclosure label. No effects on source trustworthiness were observed, and adding the text about the practice of personalized advertising did not increase PK or encourage different responses to the message.

Building on the work of Kruikemeier et al. (2016), the current study implements new dependent variables (i.e., source and message credibility), a different PMT simulation, and a different type of targeting disclosure labeling. Where Kruikemeier et al. (2016) used a “sponsored” label as their personalized advertisement condition, the current research will ask participants about their view on climate change regulations and show them a Facebook advertisement that is either for or against these regulations. As such, we try to show participants an advertisement that aligns with their views to simulate PMT. Furthermore, instead of a training condition with an explanation of micro-targeting that may take longer to read, we focus on a potential new *targeting disclosure* label that provides the receiver with more salient information in the form of a short sentence. This corresponds to the disclosure labels that Facebook itself uses and hereby fits the design of the advertisements on that platform.

Figure 1: Proposed Mediation Model



### 3 Method

The study was approved by the ethics committee of the University of Duisburg-Essen. We pre-registered this study before collecting data: <https://osf.io/nbtc4>. All participants gave informed consent before participation. Supplementary materials are publicly accessible on OSF (<https://osf.io/z4wb8>).

#### 3.1 Analysis

To test differences in the effects of a targeting disclosure label (vs. regular disclosure label and no disclosure label), we used the PROCESS macro, model 4, by Hayes (2018) using 5,000 bootstrap samples. Hypothesis 1 represents the total effects, or path c, within the mediation model. Hypothesis 2 represents path a, the first half of the indirect effects, and Hypothesis 3 represents the product of path a and path b, the indirect effects. The mediation analyses were run a total of six times, once with the targeting disclosure label group dummy as an independent variable, the regular disclosure label category as the reference category, and the no disclosure label condition as a covariate (Hayes & Preacher, 2014). The second time the mediation analysis was run with the targeting disclosure label group dummy as an independent variable, the no disclosure label category as the reference category, and the regular disclosure label category as a covariate. Furthermore, the analysis was run separately for each of the three dependent variables (i.e., source credibility, message credibility, and source trustworthiness).

### 3.2 Sample

We determined the sample size using MedPower by Kenny (2017). Based on the work of Kruikemeier et al. (2016), we assumed a small effect size. We added approximately 17% to the indicated sample size ( $N=171$ ) to arrive at a target of at least  $N=200$ .

We recruited 280 German Facebook users via the online non-probability access panel of German company Respondi AG (which, following a merger, now pertains to Bilendi & Respondi) during the period September 20–24, 2021. Thirteen respondents did not agree to our informed consent, 13 respondents were not Facebook users, and seven failed to correctly answer our attention check items. Additionally, 20 participants did not finish the experiment, leaving a sample of 227 respondents. Participant age ranged from 20 to 69 years old ( $M=43.84$ ,  $SD=14.03$ ). One hundred and twelve participants identified as female and 115 as male. Regarding education, most participants reported obtaining a secondary school certificate ( $n=90$ ), 57 reported a university entrance qualification, 39 reported a master's degree, 32 reported a bachelor's degree, 28 reported a qualifying middle school diploma, and 15 reported an advanced technical college entrance diploma, 2 indicated having no diploma, 2 indicated a doctorate, and 5 indicated that their educational background differed from all the available responses. Randomization checks showed that the experimental groups did not differ in terms of age ( $F(2, 224)=1.79$ ,  $p=0.169$ ), gender ( $\chi^2(2, N=227)=2.71$ ,  $p=.260$ ) or highest level of education ( $\chi^2(16, N=227)=18.01$ ,  $p=.320$ ).

### 3.3 Study Design

To test our hypotheses, we conducted an online experiment with a factorial between-subjects design. The experiment entailed three conditions: participants were exposed to either a regular Facebook post without any disclosure label, a regular Facebook advertisement with the disclosure label that Facebook uses on its platform (“Sponsored, paid for by ...”), or a Facebook advertisement with a more salient targeting disclosure label (“This content has been targeted at you based on your online behavior”).

To at least simulate micro-targeting, we tried to expose participants to advertisements in line with their own beliefs. To accomplish this, we asked the participants to indicate whether they agreed or disagreed with a set of statements on climate change regulations. These statements were pre-tested in a set of eight statements ( $N=100$ ) to determine which statements clearly described the view of someone who is for climate change regulations (the view of someone who wants to counter climate change and could be considered more eco-friendly, hereafter called *pro*) or against climate change regulations (the view of

someone who does not believe in climate change and thinks the regulations to counter this do not work, hereafter called *against*). Four of these pre-tested statements – those most clearly pro or against – were used in the main study.

### 3.4 Procedure

Of the ten statements used in the main study, four statements were either pro or against climate change regulations (two each). Six statements served to counter testing effects. After answering the statements, participants were randomly assigned to one of the experimental conditions. Depending on whether they agreed or disagreed with the statements, we measured their views on climate change regulations, and participants were shown a Facebook post featuring a statement that aligned with their answers to the previous statements. In this way, we tried to recreate at least a degree of (micro) targeting by showing them something a well-developed algorithm might also show them. After exposure to our stimuli, the participants answered the questions regarding our variables.

Thereafter, we asked participants to state the degree to which they found the advertisements targeted and in line with their beliefs regarding climate change regulations. Finally, we asked for demographic information (see the measures section) before debriefing and thanking the participants. The average completion time was about 8 minutes.

### 3.5 Stimuli

We created a total of six Facebook advertisements for a non-existent male politician from a non-existent political party (see Figure 2). For each experimental condition, one Facebook post was developed: a Facebook post with no disclosure label, a Facebook post with a regular disclosure label (“sponsored”), and a Facebook post with a targeting disclosure label. We developed one of each of these posts for pro participants and one of each of these posts for against participants. Other than the disclosure labels and the statement used in the advertisement, all stimuli were identical (i.e., in terms of the number of likes and number of comments). On Facebook, a regular disclosure label is normally placed in the same position as the time of posting on a non-sponsored post. The chosen posting time for the no disclosure label condition was “Yesterday, 09:32.” This was chosen because we did not have to change the date based on when the participant took part in the experiment and because the time is during office hours. We used the same locations that Facebook uses for our disclosure labels to increase the study’s external validity.

Figure 2: Examples of the Sponsored Disclosure Label, the No-Disclosure Label, and the Total Post Including the Targeting Disclosure Label as Used in the Experiment (in German)



### 3.6 Measures

All measured constructs were tested regarding their factor validity in confirmatory factor analysis. The output of this analysis and an overview of our items can be found as online supplementary materials and on OSF (<https://osf.io/z4wb8>). The factor analysis revealed a near-adequate model fit for all indicators (Alavi et al., 2020; Hair et al., 2014):  $\chi^2(98)=256.03, p<.001, \chi^2/df=2.61, CFI=0.95, TLI=0.94, RMSEA = .08, 90\% CI [.07, .10], SRMR = .04$ . (CFI=Comparative fit index; TLI=Tucker-Lewis index; RMSEA=root-mean-square error of approximation; SRMR=standardized root-mean-square residual; CI=confidence interval).

We measured the mediator PK using five items developed by Kruikemeier et al. (2016) based on Ham et al. (2015) that have been validated in the context of prior research on micro-targeting. The items (e.g., “The post feels like an ad”) were measured on a seven-point Likert scale ranging from 1 (= strongly disagree) to 7 (= strongly agree). The mean score of these items was calculated and used as a measure for PK (Cronbach’s  $\alpha = .83$ , McDonald’s  $\omega = .84$ , Average Variance Extracted = .53).

In addition, following Winter and Krämer (2014), we measured the dependent variable source credibility using three items with five-point semantic scales to establish a credibility score (ranging from, for example, “competent” to “incompetent”). Here, the mean score of these items was calculated and used as a measure for source credibility (Cronbach’s  $\alpha = .93$ , McDonald’s  $\omega = .93$ , Average Variance Extracted = .82).

Furthermore, we measured message credibility using three items developed by Appelman and Sundar (2016). Participants were asked to assess how well the adjectives “accurate,” “authentic,” and “believable” described the content they had just read using a seven-point Likert scale ranging from 1 (= describes very poorly) to 7 (= describes very well). The mean score of these items was calculated and used to measure message credibility (Cronbach’s  $\alpha = .94$ , McDonald’s  $\omega = .94$ , Average Variance Extracted = .84).

Finally, following Ohanian (1990) and, more recently, the PMT research of Kruikemeier et al. (2016), we measured source trustworthiness via four items using seven-point semantic differential scales to establish a trustworthiness score (ranging from, for example, “dishonest” to “honest”). Due to an overlap in translation into German, the item for “dependable–undependable” was not measured in this study. We have extended the scale with one item used by Winter and Krämer (2014), namely, “sincere–insincere,” which is also measured on a seven-point semantic differential scale. Again, the mean score of these items was calculated and used to measure source trustworthiness (Cronbach’s  $\alpha = .93$ , McDonald’s  $\omega = .93$ , Average Variance Extracted = .74).

As a manipulation check, we asked participants whether they recalled seeing the disclosure label (“On the Facebook post, there was a disclosure message about it being sponsored”), seeing the targeting disclosure label (“On the Facebook post, there was a disclosure message about it being tailored to my online behavior”), or not seeing any disclosure label (“There was no disclosure message on the Facebook post”). We also included several control variables and asked participants for demographic information (age, gender, and highest level of education).

Finally, the pre-registered variable *perceived targeting* was measured by two items: “I think the Facebook message was tailored to me” and “The Facebook message matched with my personal views.” These items were dichotomously measured due to a mistake in data collection. However, the variable showed no variance and was therefore omitted from the analyses.

## 4 Results

This study’s focus concerned how positioning disclosure labels above a micro-targeted political Facebook post impacts source and message credibility and source trustworthiness. Furthermore, we investigated the mediating role of PK on these direct effects. Our preliminary analyses are publicly accessible on OSF (<https://osf.io/zf83k>). The bivariate correlations, means, and standard deviations for our measured variables appear in Table 1.

Table 1: Means, Standard Deviations, and Bivariate Correlations of the Measured Constructs

Measured construct	<i>M</i> ( <i>SD</i> )	1	2	3
1 Persuasion Knowledge	4.02 (1.12)	-		
2 Source Credibility	3.11 (1.04)	-0.05	-	
3 Message Credibility	4.33 (1.51)	-0.07	0.73***	-
4 Source Trustworthiness	4.23 (1.40)	-0.07	0.82***	0.81***

Note. \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ .

According to our manipulation check, a substantial proportion of participants did not recall the correct disclosure (see Table 2). Because the assumption for normality could not be met and the sub-sample sizes for the groups (including recollection) were not close to equal, we compared the differences in the means of our groups for the mediator and the dependent variables between the participants who recalled the label correctly and the participants who did not. This involved using the Mann-Whitney U test for PK ( $U=6012$ ,  $p=.962$ ), message credibility ( $U=5638.5$ ,  $p=.405$ ), source credibility ( $U=5587.5$ ,  $p=.345$ ), and source trustworthiness ( $U=5796.5$ ,  $p=.617$ ). After comparing the groups, we found no significant differences and chose to keep our whole sample for further analyses.

Means and standard deviations appear in Table 3, and means and standard deviations for participants that correctly recalled the label can be found in the online supplementary materials.

Table 2: Results of the Manipulation Check

Condition	<i>n</i>	Recall	Percentage
Control (no disclosure)	76	54	71
Sponsored disclosure	75	18	24
Targeting disclosure	76	13	17

Table 3: Mean Scores (with Standard Deviations Between Parentheses)

Measure	No disclosure ( <i>n</i> =76)	Sponsored disclosure ( <i>n</i> =75)	Targeting disclosure ( <i>n</i> =76)
Persuasion Knowledge	4.09 (1.16)	3.94 (1.14)	4.02 (1.07)
Source Credibility	3.00 (1.06)	3.14 (1.07)	3.18 (1.00)
Message Credibility	4.35 (1.53)	4.41 (1.57)	4.23 (1.44)
Source Trustworthiness	4.21 (1.40)	4.24 (1.53)	4.23 (1.27)

Note. *N*=227. All constructs were measured on a 7-point Likert or semantic scale.

Contrary to our hypothesis ( $H_1$ ), there were only small – but non-significant – increases in ( $H_{1a}$ ) source credibility for the targeting disclosure label group compared to the regular disclosure label group ( $b=0.04$ ,  $t(224)=0.27$ ,  $p=.788$ ) and the no disclosure label group ( $b=0.18$ ,  $t(224)=0.17$ ,  $p=.276$ ) when controlling for the other experimental group. Furthermore, there were only minor – and non-significant – reductions in message credibility for the targeting disclosure label group compared to the regular disclosure label group ( $b=-0.18$ ,  $t(224)=-0.73$ ,  $p=.463$ ) and the no disclosure label group ( $b=-0.12$ ,  $t(224)=-0.48$ ,  $p=.630$ ) when controlling for the other experimental group. Additionally, there were no reductions in ( $H_{1b}$ ) source trustworthiness for the targeting disclosure label group compared to the regular disclosure label group ( $b=-0.01$ ,  $t(224)=-0.04$ ,  $p=.971$ ) and almost no increases in the targeting disclosure label compared to the no disclosure label group ( $b=0.02$ ,  $t(224)=0.10$ ,  $p=.917$ ) when controlling for the other experimental group.

Regarding our second hypothesis, we found small – but non-significant – increases in persuasion knowledge ( $H_2$ ) for the targeting disclosure label group compared to the regular disclosure label group ( $b=0.08$ ,  $t(224)=0.44$ ,  $p=.663$ ). By contrast, we observed a small – but non-significant – decrease in PK for the targeting disclosure label group compared to the no disclosure label group ( $b=-0.07$ ,  $t(224)=-0.39$ ,  $p=.697$ ) when controlling for the other experimental group.

To test the third hypothesis, we considered the mediation – or indirect effects (path a \* path b) – within the analysis conducted using PROCESS Macro, which included bootstrapping with 1,000 samples (Hayes, 2018). Contrary to what we hypothesized, but aligning with the results of our first two hypotheses, we found no results that would lead us to accept the hypothesis. Results of the mediation analyses appear online (<https://osf.io/xv7kt>). Notably, although we observed no direct effects, we recognize that this does not always mean that there are no indirect effects (Hayes, 2009, p. 413). Instead, it can indicate the possibility that an unobserved variable might have influenced the model (cf. Bullock et al., 2010, p. 551). A summary of our findings appears in Table 4.

Table 4: Summary of Findings

Hypothesis	Testing Result
$H_{1a}$ A micro-targeted political Facebook message with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will lower the perceived credibility of both the source and the message.	Rejected
$H_{1b}$ A micro-targeted political Facebook message with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will lower the perceived trustworthiness of the source.	Rejected
$H_2$ A micro-targeted political Facebook advertisement with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will activate stronger persuasion knowledge.	Rejected
$H_{3a}$ A micro-targeted political Facebook advertisement with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will activate stronger persuasion knowledge, negatively impacting source and message credibility.	Rejected
$H_{3b}$ A micro-targeted political Facebook advertisement with a targeting disclosure label (vs. a regular disclosure label or no disclosure label) will activate stronger persuasion knowledge, negatively impacting source trustworthiness.	Rejected

## 4.1 Additional Analyses

After conducting the pre-registered analyses, our results demonstrated that there were no significant effects and that none of our hypotheses could be supported. To further investigate our findings, we performed additional Bayesian analyses using JASP (JASP Team, 2021). By adding Bayesian hypothesis testing, the probability of the observed data given the null hypothesis ( $H_0$ ) is compared to the probability of the observed data given the alternative hypothesis ( $H_1$ ) (Wagenmakers, 2007). The Bayes Factor ( $BF_{01}$ ) is a ratio of these probabilities and is commonly interpreted as the weight of evidence in support of the null versus the alternative hypothesis. Our interpretation of Bayes Factors will rely on Wagenmakers et al. (2011).

For our first hypothesis, we conducted three Bayesian ANOVAs for our dependent variables with exposure to the targeting label as a fixed factor. Concerning source credibility, the analysis showed substantial evidence for  $H_0$  compared to  $H_1$  ( $BF_{01}=4.875$ ), favoring the absence of the effect of the targeting disclosure label on source credibility. Next, regarding message credibility, the analysis showed substantial evidence for  $H_0$  compared to  $H_1$  ( $BF_{01}=5.184$ ), favoring the absence of any effect of the targeting disclosure label on message credibility. Next, concerning source trustworthiness, the analysis showed substantial evidence for  $H_0$  compared to  $H_1$  ( $BF_{01}=6.529$ ), favoring the absence of the effect of the targeting disclosure label on source trustworthiness. For our second hypothesis, we also conducted a Bayesian ANOVA for our proposed mediator (PK) with exposure to the targeting label as a fixed factor. For PK, this ANOVA showed substantial evidence for  $H_0$  compared to  $H_1$  ( $BF_{01}=6.532$ ), favoring the absence of any effect of the targeting disclosure label on message credibility.

Our third hypothesis proposed a stronger mediation effect for exposure to a targeting disclosure label compared to exposure to a regular disclosure label or no exposure to any disclosure label. However, because there is substantial evidence for the absence of any effect of the targeting label on persuasion knowledge, we cannot investigate a possible mediation effect using Bayesian analysis.

## 5 Discussion

This study aimed to investigate how targeting disclosure labels on micro-targeted political advertisements on Facebook impact source credibility, message credibility, and source trustworthiness. Furthermore, we investigated the mediating role of PK on these direct effects. To contribute to the existing body of research, we incorporated source and message credibility as well as source trustworthiness as dependent variables. Additionally, we exposed participants to statements that aligned with their views in an attempt to simulate micro-targeting.

When considering the direct effects of our manipulation on our dependent variables, we found minor and non-significant differences between the three experimental groups for our dependent variables. Exposure to the targeting disclosure label led to minor increases in source credibility. As assumed, exposure to a targeting disclosure label led to small decreases in message credibility. Contradicting our expectations regarding source trustworthiness, the lowest mean was observed for the no disclosure label group, with the means for the other two groups almost identical. Ultimately, however, we rejected our first hypothesis because no significant differences were found. For our second hypothesis, supporting the findings of Kruikemeier et al. (2016), we observed no differences in the activation of PK between our experimental conditions. Additionally, our third hypothesis did not hold true: We observed no mediating effect of PK on the impact of our manipulations on source credibility, message credibility, and source trustworthiness. This can be explained by the fact that we did not observe any direct effect of those manipulations on these dependent variables. Additionally, we investigated our data using Bayesian analysis to further investigate our initial findings (Wagenmakers, 2007). In conclusion, all our analyses regarding our first two hypotheses yielded results favoring the absence of any of our proposed effects. We were not able to test our third hypothesis due to the evidence favoring the absence of any effect of our manipulations on either the proposed mediating variable or the dependent variables.

The most consistent observation throughout our findings is the fact that participants seemed not to notice the disclosure labels. This aligns with other research on disclosures that has similarly noted that not all participants recalled the disclosure messages (Evans et al., 2017; Kruikemeier et al., 2016; van Reijmersdal et al., 2021; Wojdyski & Evans, 2016). However, much of the previous work recorded higher levels of disclosure recollection than observed in the current study. Although we tried to manipulate the disclosure labels in a manner resembling Facebook's approach to achieve ecological validity, we recognize that, in hindsight, this meant we developed seemingly too-subtle disclosures. As discussed, selling advertising space on the platform is a large part of Facebook's business model, and we wanted to make sure we at least tried to design a disclosure label that could be implemented without disturbing the layout of the advertisements or posts. Our findings suggest that our targeting disclosure labels were too subtle to be recognized by participants, precluding their contribution to the recognition of advertisements or (in turn) to the transparency of targeting procedures. Furthermore, participants were exposed to the disclosure labels in an experimental setting in which they saw only one advertisement. In a real-world setting, they would be exposed to multiple posts and advertisements when browsing their Facebook timeline, making the recall of disclosure labels potentially even harder. Whatever the case, our disclosure labels were ultimately too subtle to even be recognized by participants.

## 5.1 Implications

Platforms should investigate the disclosures they use because disclosure messages are sometimes not recalled correctly or even noticed. Although our goal was not to evaluate the effectiveness of Facebook's disclosure practices, we now clearly see that it uses disclosure labels that are too subtle (cf. Boerman & Kruikemeier, 2016; Kruikemeier et al., 2016) and that do not get noticed substantially enough to be encoded in a manner that increases transparency (cf. Binford et al., 2021). Although other researchers have demonstrated promising results concerning the recall of information from disclosures, the most significant prerequisite for this is that disclosures are actually noticed in the first place (Binford et al., 2021; van Reijmersdal et al., 2016). Furthermore, the prominence of the disclosure seems important (Amazeen & Wojdyski, 2020; Boerman et al., 2015), which could imply that although platforms try to keep everything within their own corporate layout, they do not yet do enough to inform their audiences.

One of the implications of our findings concerns the processing of cues using a heuristic route, as explained in the Elaboration Likelihood Model (Petty & Cacioppo, 1986). When people notice a disclosure label, they are more likely to know that the message has been designed to persuade them, meaning that the central processing of arguments might be activated. This, in turn, can lead to a more critical evaluation of both the sender and the message. In the current setting, a user's focus might be on the Facebook post itself instead of the disclosure label used, which would also explain the low number of correct recalls observed in our study, even after using a timer to ensure participants viewed the manipulation for at least 30 seconds. Although it makes sense for users to focus on the content itself, disclosure labels are used to increase transparency and inform them about the nature of the content they consume, making it, again, important for platforms to ensure disclosures are more prominent and (therefore) more likely to be seen. Disclosures should be used to inform users instead of being used for the sake of being used.

Users are exposed to large amounts of content on SNSs. Some of this content is posts or pictures of friends and messages from groups they follow. Meanwhile, some content is advertising from companies and political parties. Although most SNS users have been exposed to persuasive messages on these platforms in the past, it can still be hard to distinguish between ads and actual user content (Wojdyski & Evans, 2016). PK helps users manage these persuasive attempts and we maintain – partly based on the current body of work regarding disclosures and their effectiveness – that using disclosure labels can give users a better chance of recognizing these attempts and better distinguishing between tailored and non-tailored advertising.

## 5.2 Limitations and Future Work

The current study's biggest limitation was participants' lack of recall of our manipulation disclosures. Although we aimed to achieve high levels of ecological validity by using Facebook's current sponsored disclosure format and developing targeting disclosure messages to resemble it, we recognize that we might have overamplified that idea to the point that our disclosures were too subtle to be noticed by our participants. Therefore, the current body of work would benefit substantially from more research concerning the design of disclosures, especially the factors that can improve recognition of disclosures. At the same time, we recognize that most studies measure recall or perform manipulation checks using self-report approaches rather than including a measurement directly after exposure to the stimuli, potentially causing users to check the wrong option. This is possible for participants in all experimental conditions. One solution would be to ask participants about the disclosures on the post directly after exposure and treat the recall question as a measure and not just a manipulation check. Another option would be to use eye-tracking. Using eye-tracking would enable researchers to not only measure the visual fixation on the disclosures but – if combined with the recall question – also assess potential differences in gazing behavior and recall. This is because looking at a disclosure might not be the same as internalizing it and understanding its meaning.

Additionally, we also recognize that we did not include the information that the targeted post was an advertisement in the disclosure for our targeting condition. We wanted to keep the disclosure short, so we did not include this information. In hindsight, it is possible that there was no difference between our experimental groups in terms of PK because we did not provide that information to our participants. Furthermore, we recognize that content being targeted does not also necessarily mean that it is sponsored. For instance, Facebook also shows users content that they are more likely to engage with based on their past interests and interactions. However, we did state that the content was sponsored in the “sponsored” condition, but we did not find any effects in that group compared to the control condition without a disclosure.

Meanwhile, although we measured the credibility of the message and the source, we only considered the trustworthiness of the source, ignoring the trustworthiness of the message. This might represent an opportunity for future work. In hindsight, if we had included a measure for this, we would have been able to provide a clearer overview of potential differences in perceptions between the message and the source based on different disclosure labels.

This study was conducted around the time of the 2021 German general elections. As such, the opinions of participants on the statements used for the micro-targeting simulation might have been more explicit than usual. Nonetheless, because the sample comprised only German citizens, these potential effects would have been identical across the whole sample. Within the manipulation we constructed, we made use of a made-up German political party

(“Deutschland zusammen”). We did this to ensure that no participants were influenced by previous exposure to an advertisement for a real political party during the period around the general elections. We did this to protect our participants from any possible persuasion during our experiment, but we also recognize that this possibly influenced our findings. Although participants viewed the Facebook post for a minimum of 30 seconds, a novelty effect might have taken place. We tried to match the post to the views of the participant, which might have interested them in the political party and made them more focused on the name and profile picture of the party, undermining the possibility of them paying attention to the manipulated disclosure labels. When considering the mean score for source credibility, it is overall lower than the other measured constructs – this might imply that people were thinking too much about the fake political party and questioning its credibility, instead of paying attention to the whole Facebook post.

Because PMT in its current form is a rather new phenomenon, the actual practices of consulting firms, political parties, and SNSs remain subject to debate. Furthermore, the influence of PMT on specific political instances (e.g., Trump’s campaign, the Brexit campaign) continues to be questioned. These inquiries are indeed valid, and it seems critical to try to inform the public about the practices that allegedly take place on the SNSs where they consume a considerable volume of information, especially given the potential threat to democracy that PMT might represent. Additionally, we maintain that – because they have been found to predict voting behavior – credibility and trustworthiness are important constructs for investigations into PMT. That is, although this study’s findings might differ from our expectations, we remain convinced that it is necessary to develop countermeasures that can be implemented in ways that are acceptable for SNSs, and we encourage future researchers to keep that in mind while exploring novel implementations.

## 6 Conclusion

Considering as a mediating factor PK, a mechanism that has been studied in depth in the context of consumer research, this work has studied the effects on credibility and trustworthiness of informing users about the targeting practices taking place on SNSs. Contrary to expectations, we did not observe any effects of our manipulations on credibility or trustworthiness. Supporting extant research, a large part of our sample did not recall seeing any disclosure label, regardless of the manipulation they were exposed to. Although this is an interesting result, it introduced complications around measuring the actual effects of the disclosure labels. This does represent a problem, but we would like to emphasize that it is also the largest takeaway from this study, especially given that one disclosure label resembled Facebook's standard practice and another disclosure label moved beyond this by including additional information and making the disclosure more salient. Platforms use disclosures as a transparency measure to inform users about advertising practices or to at least try to or appear to be trying to. However, if users do not notice disclosure labels in the current format, this might be perceived to be an insufficient effort, window-dressing, or even empty transparency.

## References

- Aguirre, E., Mahr, D., Grewal, D., de Ruyter, K., & Wetzels, M. (2015). Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of Retailing*, 91(1), 34–49. <https://doi.org/10.1016/j.jretai.2014.09.005>
- Alavi, M., Visentin, D. C., Thapa, D. K., Hunt, G. E., Watson, R., & Cleary, M. (2020). Chi-square for model fit in confirmatory factor analysis. *Journal of Advanced Nursing*, 76(9), 2209–2211. <https://doi.org/10.1111/jan.14399>
- Amazeen, M. A., & Wojdyski, B. W. (2020). The effects of disclosure format on native advertising recognition and audience perceptions of legacy and online news publishers. *Journalism*, 21(12), 1965–1984. <https://doi.org/10.1177/1464884918754829>
- Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism and Mass Communication Quarterly*, 93(1), 59–79. <https://doi.org/10.1177/1077699015606057>

- Barbu, O. (2014). Advertising, Microtargeting and Social Media. *Procedia – Social and Behavioral Sciences*, 163, 44–49. <https://doi.org/10.1016/j.sbspro.2014.12.284>
- Barocas, S. (2012). The price of precision: Voter microtargeting and its potential harms to the democratic process. *Proceedings of the First Edition Workshop on Politics, Elections and Data – PLEAD '12*, 31. <https://doi.org/10.1145/2389661.2389671>
- Ben-David, A. (2020). Counter-archiving Facebook. *European Journal of Communication*, 35(3), 249–264. <https://doi.org/10.1177/0267323120922069>
- Binford, M. T., Wojdyski, B. W., Lee, Y.-I., Sun, S., & Briscoe, A. (2021). Invisible transparency: Visual attention to disclosures and source recognition in Facebook political advertising. *Journal of Information Technology & Politics*, 18(1), 70–83. <https://doi.org/10.1080/19331681.2020.1805388>
- Bodó, B., Helberger, N., & De Vreese, C. H. (2017). Political micro-targeting: A manchurian candidate or just a dark horse? *Internet Policy Review*, 6(4). <https://doi.org/10.14763/2017.4.776>
- Boerman, S. C., & Kruikemeier, S. (2016). Consumer responses to promoted tweets sent by brands and political parties. *Computers in Human Behavior*, 65, 285–294. <https://doi.org/10.1016/j.chb.2016.08.033>
- Boerman, S. C., & van Reijmersdal, E. A. (2016). Informing Consumers about “Hidden” Advertising: A Literature Review of the Effects of Disclosing Sponsored Content. In P. De Pelsmacker (Ed.), *Advertising in New Formats and Media* (pp. 115–146). Emerald Group Publishing Limited. <https://doi.org/10.1108/978-1-78560-313-620151005>
- Boerman, S. C., van Reijmersdal, E. A., & Neijens, P. C. (2012). Sponsorship Disclosure: Effects of Duration on Persuasion Knowledge and Brand Responses. *Journal of Communication*, 62(6), 1047–1064. <https://doi.org/10.1111/j.1460-2466.2012.01677.x>
- Boerman, S. C., van Reijmersdal, E. A., & Neijens, P. C. (2015). Using Eye Tracking to Understand the Effects of Brand Placement Disclosure Types in Television Programs. *Journal of Advertising*, 44(3), 196–207. <https://doi.org/10.1080/00913367.2014.967423>
- Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what’s the mechanism? (Don’t expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550–558. <https://doi.org/10.1037/a0018933>
- Cadwalladr, C. (2018, March 18). ‘I made Steve Bannon’s psychological warfare tool’: Meet the data war whistleblower. *The Guardian*. <https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump>

- Campbell, M. (1995). When Attention-Getting Advertising Tactics Elicit Consumer Inferences of Manipulative Intent: The Importance of Balancing Benefits and Investments. *Journal of Consumer Psychology*, 4(3), 225–254. [https://doi.org/10.1207/s15327663jcp0403\\_02](https://doi.org/10.1207/s15327663jcp0403_02)
- Carr, C. T., & Hayes, R. A. (2014). The Effect of Disclosure of Third-Party Influence on an Opinion Leader's Credibility and Electronic Word of Mouth in Two-Step Flow. *Journal of Interactive Advertising*, 14(1), 38–50. <https://doi.org/10.1080/15252019.2014.909296>
- Coppock, A., Green, D. P., & Porter, E. (2022). Does digital advertising affect vote choice? Evidence from a randomized field experiment. *Research & Politics*, 9(1), 205316802210769. <https://doi.org/10.1177/20531680221076901>
- De Keyzer, F., van Noort, G., & Kruikemeier, S. (2022). GOING TOO FAR? HOW CONSUMERS RESPOND TO PERSONALIZED ADVERTISING FROM DIFFERENT SOURCES. *Journal of Electronic Commerce Research*, 23(3), 22.
- Deng, X., Li, M., & Suh, A. (2020). Recommendation or advertisement? The influence of advertising-disclosure language with pictorial types on influencer credibility and consumers' brand attitudes. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12427 LNCS, 234–248. [https://doi.org/10.1007/978-3-030-60152-2\\_19](https://doi.org/10.1007/978-3-030-60152-2_19)
- Dobber, T., Trilling, D., Helberger, N., & de Vreese, C. H. (2017). Two crates of beer and 40 pizzas: The adoption of innovative political behavioural targeting techniques. *Internet Policy Review*, 6(4). <https://doi.org/10.14763/2017.4.777>
- Dommett, K. (2020). Regulating Digital Campaigning: The Need for Precision in Calls for Transparency. *Policy and Internet*, 12(4), 432–449. <https://doi.org/10.1002/poi3.234>
- Elsawah, M., & Howard, P. N. (2020). *Tunisia-memo-English.pdf* [PDF]. The Challenges of Monitoring Social Media in the Arab World: The Case of the 2019 Tunisian Elections. <https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/03/Tunisia-memo-English.pdf>
- Endres, K., & Kelly, K. J. (2018). Does microtargeting matter? Campaign contact strategies and young voters. *Journal of Elections, Public Opinion and Parties*, 28(1), 1–18. <https://doi.org/10.1080/17457289.2017.1378222>
- European Commission. (2022). *REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL* on a Single Market For Digital Services (Digital Services Act) and amending Directive. <https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:825:FIN>

- Evans, N. J., Phua, J., Lim, J., & Jun, H. (2017). Disclosing Instagram Influencer Advertising: The Effects of Disclosure Language on Advertising Recognition, Attitudes, and Behavioral Intent. *Journal of Interactive Advertising*, 17(2), 138–149.  
<https://doi.org/10.1080/15252019.2017.1366885>
- Friestad, M., & Wright, P. (1994). *The Persuasion Knowledge Model: How People Cope with Persuasion Attempts*. <https://academic.oup.com/jcr/article/21/1/1/1853712>
- Gandy, O. H. (2000). Dividing Practices: Segmentation and Targeting in the Emerging Public Sphere. In W. L. Bennett & R. M. Entman (Eds.), *Mediated Politics* (1st ed., pp. 141–159). Cambridge University Press.  
<https://doi.org/10.1017/CBO9780511613852.008>
- Ghosh, D. (2018, October 4). *What is microtargeting and what is it doing in our politics?* Mozilla. <https://blog.mozilla.org/internetcitizen/2018/10/04/microtargeting-dipayan-ghosh/>
- Giasson, T., Le Bars, G., & Dubois, P. (2019). Is Social Media Transforming Canadian Electioneering? Hybridity and Online Partisan Strategies in the 2012 Quebec Election. *Canadian Journal of Political Science*, 52(2), 323–341. <https://doi.org/10.1017/S0008423918000902>
- Gorton, W. A. (2016). Manipulating citizens: How political campaigns' use of behavioral social science harms democracy. *New Political Science*, 38(1), 61–80. <https://doi.org/10.1080/07393148.2015.1125119>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7. ed., Pearson new internat. ed). Pearson.
- Ham, C. D., Nelson, M. R., & Das, S. (2015). *How to measure persuasion knowledge*. *International Journal of Advertising*, 34(1), 17–53.  
<https://doi.org/10.1080/02650487.2014.994730>
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium. *Communication Monographs*, 76(4), 408–420.  
<https://doi.org/10.1080/03637750903310360>
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (Second edition). Guilford Press.
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67(3), 451–470.  
<https://doi.org/10.1111/bmsp.12028>
- Hetherington, M. J. (1999). The Effect of Political Trust on the Presidential Vote, 1968–96. *American Political Science Review*, 93(2), 311–326.  
<https://doi.org/10.2307/2585398>

- Housholder, E. E., & LaMarre, H. L. (2014). Facebook Politics: Toward a Process Model for Achieving Political Source Credibility Through Social Media. *Journal of Information Technology & Politics*, 11(4), 368–382. <https://doi.org/10.1080/19331681.2014.951753>
- Jamieson, K. H. (2013). Messages, micro-targeting, and new media technologies. *Forum (Germany)*, 11(3), 429–435. <https://doi.org/10.1515/for-2013-0052>
- JASP Team. (2021). *JASP* (0.14.1).
- Jung, A. R., & Heo, J. (2019). Ad Disclosure vs. Ad Recognition: How Persuasion Knowledge Influences Native Advertising Evaluation. *Journal of Interactive Advertising*, 19(1), 1–14. <https://doi.org/10.1080/15252019.2018.1520661>
- Kaiser, B. (2019). *Targeted: The Cambridge Analytica whistleblower's inside story of how big data, Trump, and Facebook broke democracy and how it can happen again*.
- Kenny, D. A. (2017). *MedPower: An interactive tool for the estimation of power in tests of mediation*. <https://davidakenny.shinyapps.io/MedPower/>
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15), 5802–5805. <https://doi.org/10.1073/pnas.1218772110>
- Kruikemeier, S., Sezgin, M., & Boerman, S. C. (2016). Political Microtargeting: Relationship between Personalized Advertising on Facebook and Voters' Responses. *Cyberpsychology, Behavior, and Social Networking*, 19(6), 367–372. <https://doi.org/10.1089/cyber.2015.0652>
- Leerssen, P., Ausloos, J., Zarouali, B., Helberger, N., & de Vreese, C. H. (2019). Platform ad archives: Promises and pitfalls. *Internet Policy Review*, 8(4), 1–21. <https://doi.org/10.14763/2019.4.1421>
- Liberini, F., Redoano, M., Russo, A., Cuevas, Á., & Cuevas, R. (2020). Politics in the Facebook Era – Evidence from the 2016 US Presidential Elections. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3584086>
- Madsen, J. K. (2019). *The psychology of micro-targeted election campaigns*. Palgrave Macmillan.
- Main, K., Dahl, D., & Darke, P. (2007). Deliberative and Automatic Bases of Suspicion: Empirical Evidence of the Sinister Attribution Error. *Journal of Consumer Psychology*, 17(1), 59–69. [https://doi.org/10.1207/s15327663jcp1701\\_9](https://doi.org/10.1207/s15327663jcp1701_9)

- Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America*, 114(48), 12714–12719. <https://doi.org/10.1073/pnas.1710966114>
- Murray, G. R., & Scime, A. (2010). Microtargeting and electorate segmentation: Data mining the American National Election Studies. *Journal of Political Marketing*, 9(3), 143–166. <https://doi.org/10.1080/15377857.2010.497732>
- Ohanian, R. (1990). Construction and Validation of a Scale to Measure Celebrity Endorsers' Perceived Expertise, Trustworthiness, and Attractiveness. *Journal of Advertising*, 19(3), 39–52.
- Ortega, A. L. (n.d.). *Are microtargeted campaign messages more negative and diverse? An analysis of Facebook ads in European election campaigns*.
- Papakyriakopoulos, O., Hegelich, S., Shahrezaye, M., & Serrano, J. C. M. (2018). Social media and microtargeting: Political data processing and the consequences for Germany. *Big Data & Society*, 5(2), 2053951718811844. <https://doi.org/10.1177/2053951718811844>
- Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of Persuasion. *Advances in Experimental Social Psychology*, 19, 123–205. [https://doi.org/10.1016/S0065-2601\(08\)60214-2](https://doi.org/10.1016/S0065-2601(08)60214-2)
- Segijn, C. M., & van Ooijen, I. (2022). Differences in consumer knowledge and perceptions of personalized advertising: Comparing online behavioural advertising and synced advertising. *Journal of Marketing Communications*, 28(2), 207–226. <https://doi.org/10.1080/13527266.2020.1857297>
- Stubb, C., & Colliander, J. (2019). “This is not sponsored content” – The effects of impartiality disclosure and e-commerce landing pages on consumer responses to social media influencer posts. *Computers in Human Behavior*, 98, 210–222. <https://doi.org/10.1016/j.chb.2019.04.024>
- Van Der Goot, M. J., Van Reijmersdal, E. A., & Zandbergen, S. K. P. (2021). Sponsorship Disclosures in Online Sponsored Content: Practitioners' Considerations. *Journal of Media Ethics*, 36(3), 154–169. <https://doi.org/10.1080/23736992.2021.1935962>
- van Reijmersdal, E. A., Fransen, M. L., van Noort, G., Oprea, S. J., Vandenberg, L., Reusch, S., van Lieshout, F., & Boerman, S. C. (2016). Effects of Disclosing Sponsored Content in Blogs: How the Use of Resistance Strategies Mediates Effects on Persuasion. *American Behavioral Scientist*, 60(12), 1458–1474. <https://doi.org/10.1177/0002764216660141>

- van Reijmersdal, E. A., Rozendaal, E., Hudders, L., Vanwesenbeeck, I., Cauberghe, V., & van Berlo, Z. M. C. (2020). Effects of Disclosing Influencer Marketing in Videos: An Eye Tracking Study Among Children in Early Adolescence. *Journal of Interactive Marketing*, 49, 94–106. <https://doi.org/10.1016/j.intmar.2019.09.001>
- van Reijmersdal, E. A. van, Oprea, S. J., & Cartwright, R. F. (2021). Brand in focus: Activating adolescents' persuasion knowledge using disclosures for embedded advertising in music videos. *Communications*. <https://doi.org/10.1515/commun-2019-0168>
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, 14(5), 779–804. <https://doi.org/10.3758/BF03194105>
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426–432. <https://doi.org/10.1037/a0022790>
- Wentzel, D., Tomczak, T., & Herrmann, A. (2010). The moderating effect of manipulative intent and cognitive resources on the evaluation of narrative ads. *Psychology and Marketing*, 27(5), 510–530. <https://doi.org/10.1002/mar.20341>
- Wilson, D. G. (2017). The ethics of automated behavioral microtargeting. *AI Matters*, 3(3), 56–64. <https://doi.org/10.1145/3137574.3139451>
- Winter, S., & Krämer, N. C. (2014). A question of credibility – Effects of source cues and recommendations on information selection on news sites and blogs. *Communications*, 39(4), 435–456. <https://doi.org/10.1515/commun-2014-0020>
- Winter, S., Maslowska, E., & Vos, A. L. (2021). The effects of trait-based personalization in social media advertising. *Computers in Human Behavior*, 114. <https://doi.org/10.1016/j.chb.2020.106525>
- Wojdyski, B. W., & Evans, N. J. (2016). Going Native: Effects of Disclosure Position and Language on the Recognition and Evaluation of Online. *Journal of Advertising*, 45(2), 157–168. <https://doi.org/10.1080/00913367.2015.1115380>
- Wylie, C. (2019). *Mindf\*ck: Cambridge Analytica and the plot to break America* (First edition). Random House.
- Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How much can behavioral targeting help online advertising? *Proceedings of the 18th International Conference on World Wide Web – WWW '09*, 261. <https://doi.org/10.1145/1526709.1526745>

Zarouali, B., Dobber, T., De Pauw, G., & de Vreese, C. (2020). Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research*.

<https://doi.org/10.1177/0093650220961965>

Zuiderveen Borgesius, F. J., Möller, J., Kruikemeier, S., Fathaigh, R., Irion, K., Dobber, T., Bodo, B., & de Vreese, C. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, 14(1), 82–96. <https://doi.org/10.18352/ulr.420>

## Appendix

### Definitions of the variables

Variable	Definition	References
Source credibility	Indication if the sender is able to provide valid statements on a topic, according to the receiver.	Winter and Krämer, 2014
Message credibility	Indication of an individual's judgement of the veracity of the content of communication.	Appelman and Sundar, 2016
Source trustworthiness	Indication of the receivers perception of honesty of the communicator.	Kruikemeier et al., 2016
Persuasion knowledge	Indicator of receivers' personal knowledge and beliefs about the motives and tactics of an advertisement and the sender of this advertisement.	Friestad and Wright, 1994

### Operationalization of the measures

Variable	Items	Scale	References
Source credibility	competent – incompetent experienced – not experienced qualified – non-qualified	5-point semantic	Winter and Krämer, 2014
Message credibility	accurate authentic believable	7-point Likert	Appelman and Sundar, 2016
Source trustworthiness	undependable – dependable <sup>a</sup> dishonest – honest <sup>a</sup> selfish – unselfish <sup>a</sup> unreliable – reliable <sup>a</sup> untrustworthy – trustworthy <sup>a</sup> sincere – insincere <sup>b</sup> The post feels like an ad	7-point semantic	<sup>a</sup> Kruikemeier et al., 2016 <sup>b</sup> Winter and Krämer, 2014
Persuasion knowledge	The post promotes the sender Sender paid to post this message The post of the sender is an ad The post is sponsored by sender	7-point Likert	Kruikemeier et al., 2016

Means and standard deviations for correct recallment

Measure	No disclosure ( <i>n</i> =54)	Sponsored disclosure ( <i>n</i> =18)	Targeting disclosure ( <i>n</i> =13)
Persuasion Knowledge	3.88 (1.06)	4.74 (1.01)	4.18 (1.13)
Source Credibility	2.93 (1.01)	3.19 (0.97)	3.23 (0.80)
Message Credibility	4.15 (1.55)	4.39 (1.67)	4.26 (0.64)
Source Trustworthiness	4.14 (1.36)	4.22 (1.51)	4.17 (0.57)

*Note.* *N*=151. All constructs were measured on a 7-point Likert or semantic scale.

Results of the mediation analyses: The indirect effects of targeting disclosure labels on the credibility of the source and message and source trustworthiness through persuasion knowledge

Variable	Indirect effect	SE	95% BCBCI	
			LL	UL
Targeting disclosure vs. regular disclosure				
Source credibility through PK	0.00	.02	-0.06	0.03
Message credibility through PK	-0.01	.03	-0.08	0.04
Source trustworthiness through PK	-0.01	.03	-0.08	0.04
Targeting disclosure vs. no-disclosure				
Source credibility through PK	0.00	.02	-0.03	0.05
Message credibility through PK	0.01	.03	-0.04	0.08
Source trustworthiness through PK	0.01	.03	-0.04	0.07

*Note.* BCBCI, bias-corrected bootstrap confidence interval; PK, persuasion knowledge; SE, standard error; LL, lower limit; UL, upper limit

## Acknowledgements

The authors thank Nur Efsan Cetinkaya for helping with the questionnaire and translations.

---

**Date received:** June 2022

**Date accepted:** June 2023