Health Communication

ISSN: 1041-0236 (Print) 1532-7027 (Online) Journal homepage: http://www.tandfonline.com/loi/hhth20

The Use of Recommendations on Physician Rating Websites: The Number of Raters Makes the Difference When Adjusting Decisions

Guillermo Carbonell, Dar Meshi & Matthias Brand

To cite this article: Guillermo Carbonell, Dar Meshi & Matthias Brand (2018): The Use of Recommendations on Physician Rating Websites: The Number of Raters Makes the Difference When Adjusting Decisions, Health Communication, DOI: 10.1080/10410236.2018.1517636

To link to this article: https://doi.org/10.1080/10410236.2018.1517636

Published online: 17 Sep 2018.

HEALTH COMMUNICATION
https://doi.org/10.1080/10410236.2018.1517636

The Use of Recommendations on Physician Rating Websites: The Number of Raters Makes the Difference When Adjusting Decisions

Guillermo Carbonell[a], Dar Meshi[b], and Matthias Brand[a,c,d]

[a]General Psychology: Cognition, University of Duisburg-Essen; [b]Department of Advertising and Public Relations, Michigan State University; [c]General Psychology: Cognition, Center for Behavioral Addiction Research (CeBAR), University Duisburg-Essen; [d]Erwin L. Hahn Institute for Magnetic Resonance Imaging, UNESCO Weltkulturerbe Zollverein

ABSTRACT

Physician rating websites allow users to check physicians' profiles, write reviews, or rate their performance. The opinion of other users regarding a physician can affect our decision to visit her/him. To investigate the specific role of the number of users rating a physician when choosing a physician with support of these platforms, we used a Judge-Advisor System in which participants answered their likelihood to visit a physician before and after seeing the recommendations of others. Within the experiment, three conditions were presented: high and low number of reviewers recommending a physician, and no recommendations. We found that the participants' likelihood to visit a physician varied with respect to the displayed physician characteristics on the platform. Importantly, after the recommendation of others was presented, participants' likelihood to visit the physician changed significantly. The participants' adjusted response was significantly closer to the recommendation coming from a higher number of users, which indicate that this online, social media cue influences our decision to visit physicians. Comments and ratings on physician ratings are generally positive, but we show that negative ratings have a direct negative influence in the decision to visit a physician. We suggest administrators of these platforms to pay special attention to the content that users upload.

Introduction

When making decisions, people can rely on the advice of others they trust. Social media plays an important role in these kinds of situations, as these online platforms allow users to check the descriptions and experiences of others regarding the acquisition of products or services. With the help of recommendations provided by others in similar situations, users can find support or discouragement when making decisions (Walther, Liang, Ganster, Wohn, & Emington, 2012). The current study uses an advice-taking paradigm commonly used in cognitive research to understand how the number of users rating a physician on a website influences their inclination to visit the physician.

Physician rating websites are online platforms with user-generated content, where users can find information about physicians and choose the one they find the most suitable for visiting. These platforms have gained in popularity over time (Emmert, Sauter, Jablonski, Sander, & Taheri-Zadeh, 2017; Gao, McCullough, Agarwal, & Jha, 2012), indicating that people are not only using these sites to choose a suitable physician, but that this type of social media platform is durable and will persist in the future. Therefore, it is important to understand whether and how these internet platforms influence users. For example, the growth of these platforms in recent years has generated many critiques from physicians. Patel, Cain, Neailey, and Hooberman (2015) showed that 17 out of 20 physicians interviewed in England were concerned about the feedback that patients post on these platforms, questioning the validity of this feedback and its potential impact. More recently, Daskivich et al. (2018) found that online ratings of physicians do not predict their actual performance. Yet, in an experimental study, Carbonell and Brand (2018) found that comments and ratings are very important features for users of these platforms when choosing a physician online. On the bright side, some studies show that the majority of ratings and reviews posted on physician rating websites are positive (Emmert, Meier, Heider, Dürr, & Sander, 2014; Kadry, Chu, Kadry, Gammas, & Macario, 2011; Lagu, Hannon, Rothberg, & Lindenauer, 2010).

Despite all the polemic around physician rating websites, the presence and use of these platforms is a current reality. Nowadays, it is normal to research, find, and choose among different products or services on the Internet, and healthcare is no exception. For this reason, it is important to understand the role of previous patients' ratings in influencing other potential patients and how these potential patients interpret this information to make subsequent decisions. Research in advice taking might be helpful to understand how patients' ratings influence the decision whether to visit a certain physician or not.

CONTACT Matthias Brand matthias.brand@uni-due.de http://www.uni-due.de/kognitionspsychologie/ General Psychology: Cognition and Center for Behavioral Addiction Research (CeBAR), University Duisburg-Essen, Forsthausweg 2, Duisburg D-47057, Germany.

Advice taking

Researchers in the fields of cognitive psychology, behavioral economics, and neuroscience have studied advice taking for many years (see Brehmer & Hagafors, 1986 for the first published paper). Many of these studies use the Judge-Advisor System (JAS, from this point on). The JAS is an advice-taking task, in which participants have to make a decision at two different times: before and after receiving advice. For instance, participants can be asked what the distance in km between New York and Paris is, to which they have to answer according to their knowledge. After answering, participants next receive advice and have the opportunity to change their first estimate. This task is used to assess the use and/or discount of advice, considering different components, such as the judge (the one who makes the decision), the advisor, the situation, and possible consequences (Bonaccio & Dalal, 2006). For instance, some studies show that in the presence of rewards, judges and advisors increase their accuracy (Sniezek, Schrah, & Dalal, 2004). Furthermore, judges are perceived as more competent when they decide to seek advice (Brooks, Gino, & Schweitzer, 2015).

Expertise and trustworthiness of an advisor have been primary variables under investigation in the advice-taking literature. Several studies indicate that individuals use the recommendations provided by confident advisors and experts more frequently (Harvey & Fischer, 1997; Önkal, Gönül, Goodwin, Thomson, & Öz, 2017; Reyt, Wiesenfeld, & Trope, 2016; Swol & Sniezek, 2005). Meshi, Biele, Korn, and Heekeren (2012) used the JAS and functional magnetic resonance imaging (fMRI) to observe the brain activation of participants when they received advice from experts and novices. On a behavioral level, they confirmed that participants usually utilize more the advice coming from experts than from novices. With regard to the neuroimaging, reward-sensitive areas (i.e., ventral striatum) were more active when participants discovered that the advice would come from an expert, even before they saw the actual advice. This finding demonstrates that judges compute the value of advice as a function of its source, in this case experts or novices. This computation, and the subsequent behavioral influence of advice, could be similar with respect to the number of people recommending (or not recommending) a physician.

One of the core characteristics of physician rating websites is that users (usually former patients) write reviews about the performance of a physician they visited. This information can be used by others who are searching for a physician. However, little is known about the expertise or trustworthiness of the reviewer. As summarized earlier, advice-taking research has shown that expert advice is more frequently used (Harvey & Fischer, 1997; Önkal et al., 2017; Reyt et al., 2016; Swol & Sniezek, 2005) and that it is more intrinsically rewarding than novice advice (Meshi et al., 2012). Yet, on a typical physician rating website, there is no information indicating the expertise of reviewers and raters, just patients or users of the platform. There are, however, some social media cues that serve as a warranting signal that indicates that some information is more credible than other information; this is termed the "warranting principle." These social media cues might function similarly to the expertise level of an advisor in an advice-taking process.

Number of recommendations as a warranting signal

Walther and Parks (2002) pointed out that certain cues on one's online profile provide credibility. They termed this the "warranting principle." The idea is that on the Internet, people are often anonymous and users do not know if others really are who they say they are. Importantly, key information cannot be manipulated, and this information provides valuable cues to users for judging the credibility of the content they see online. Walther, Van Der Heide, Hamel, and Shulman (2009) explain this phenomenon as follows: "...in warranting terms, comments provided to Person B about Person A should be more valuable to B if they come from or are corroborated by another member of A's social network (a testimonial) than if they come from Person A directly (a disclosure)" (Walther et al., 2009, p. 232). Other investigations have elaborated and confirmed this phenomenon by using different experimental approaches, such as simulating social media platforms such as Yelp (DeAndrea, Van Der Heide, Vendemia, & Vang, 2015), Facebook (Tong, Van Der Heide, Langwell, & Walther, 2008), or LinkedIn (Rui, 2017).

In regard to physician rating websites, the number of users rating a physician is a factor that can account for the credibility of the recommendation (Grabner-Kräuter & Waiguny, 2015). For example, Grabner-Kräuter and Waiguny (2015) showed that a positive attitude toward physicians is related to a higher number of reviews on their profile. In another experimental study using a choice-based conjoint design, Carbonell and Brand (2017) found that comments and ratings are primary factors on these platforms, as these are the two most important features (along with the availability of the physician) for users when they are asked to choose one among four physicians. Similarly, De Langhe, Fernbach, and Lichtenstein (2016) showed that product ratings on the Amazon website do not converge with consumer reports of experts, showing that user ratings might influence consumers to make decisions that correspond to the actual quality of a product.

The number of users providing ratings is not only important for physician rating websites, but for other social media platforms as well (Lin, Spence, & Lachlan, 2016). In a study using a different platform with user-generated content, Flanagin and Metzger (2013) performed an experiment in which they simulated a movie rating website. The movie ratings were presented in different conditions, such as the source, which could be user-generated or expert-generated, and the number of ratings. Participants were asked about: the perceived credibility of the ratings; their reliance on the ratings (if they would base the decision whether to see the movie on the rating); their confidence on the ratings (do the ratings reflect the quality of the movie?); and behavioral intentions (how likely they would be to see the movie). They observed that the aggregated opinions of non-experts regarding a product or service are important, because normally, large numbers of ratings cannot be manipulated by third parties (e.g., producers interested in the success of the movie) and are perceived as more credible (Flanagin & Metzger, 2013). Furthermore, the aggregated ratings of non-experts were perceived as less susceptible to bias, as it is the majority speaking for the quality of the movie. Therefore, the number of users rating a service or product serve as a cue that shows the user that the provided information is credible.

Studies that investigate reviews and ratings are mainly focused on rating credibility. Nevertheless, understanding credibility on social media might not be enough to observe to what extent user-generated content influences us to acquire a product or service. Even though the studies mentioned above also inquire about a potential decision, these studies are not able to show how user-generated content changes the participant's opinion toward the physician or the movie. For instance, Grabner-Kräuter and Waiguny (2015), asked participants about their attitude toward a phyisician after seeing reviews and comments, and Flanagin and Metzger (2013) asked participants about their behavioral intentions, that is, how likely they would see the movie in question after seeing the ratings and reviews in the different conditions. However, these studies do not investigate whether or not the participants changed their opinion after seeing a recommendation, because no baseline measure of the participants' opinion was included in these studies. This measure is necessary to understand how user-generated content can influence our decision to visit a physician. With this in mind, we assessed the degree of influence of physician recommendations by comparing participants' likelihood to visit a physician before and after seeing the recommendations of others, and also asking them how confident they were in their decision.

Influence of online recommendations in users' decisions

In the current study, we combined methodology used in advice-taking research with findings from communication research that suggest that the degree of influence might depend on the number of users providing ratings. We aimed to examine how, and to what extent, the number of raters on a physician rating website influences the likelihood of choosing a physician. We achieved this by using an adapted version of the JAS. Participants had to answer their likelihood to visit a physician before and after seeing the recommendation of others. Within the experiment, three conditions were presented: high and low number of recommendations for a physician, and no recommendations. Considering this, the current study aims to test the following five hypotheses:

- H1: The user's first estimate (estimate means the likelihood—in percent—to visit a physician) is based on the physician's characteristics displayed on the physician rating website.
- H2: There is a significant difference between the likelihood of a user visiting a physician before and after seeing the recommendation for a physician (irrespective of the number of ratings).
- H3: There is a significant difference in the likelihood to visit a physician in response to the number of ratings provided for a physician. In other words, the difference between users' first and second estimate should be higher when the number of ratings for a physician is high.
- H4: The users' second estimate is significantly closer to the recommendation when the number of physician ratings is high.
- H5: There is a significant difference in the confidence of users' second estimate among the three conditions. Confidence ratings are significantly higher when the number of physician ratings is high.

Methods

Participants

One hundred thirty-one subjects participated in this study. From these, one was excluded because of a technical error when exporting the collected data, resulting in the final sample of n = 130 (age: M = 25.72, SD = 9.79 years), of which 91 were women and 38 were men (one participant did not provide gender information). Subjects received €15 or course credit for participating. All subjects signed an informed consent at the start of the study, which lasted about one and a half hours. The study was approved by the local ethics committee.

Judge-Advisor System

We used a JAS paradigm to simulate the experience of visiting physician rating websites. Participants were asked to imagine that they have been experiencing several health-related symptoms over the last few weeks, namely an elevated heart rate, sweaty hands, and dizziness. The task started when participants were presented with a description of a physician on a computer screen (see below for details of this description) and they had to report on a scale from 0 to 100% how probable it was for them to visit this certain physician. After this, in the second display, subjects rated on a 10-point Likert scale, how confident they were about their response. In the third display, they were presented with the number of users who rated this physician, which could be low (1–20 users), high (250–350 users), or none (control condition). In the latter case, participants saw on the screen the message: "No ratings available." On the same screen, participants saw the percentage (from 0 to 100%) of former patients who recommended the physician. On the next two screens, participants had to report again the likelihood of visiting the physician and how confident they were about their second estimate. After this, a new trial started and a new physician was presented (see Figure 1). The entire task consisted of 120 trials, with 40 trials of each condition (low, high, and control).

Regarding the physician description, participants observed four attributes, which are commonly found on physician rating websites: *Years of experience*, with the levels 2, 4, 6, or 8 years; *Specialty*, with the levels general medicine, internist, or cardiologist; *Distance in km*, with the levels 1, 2, 4, or 8 km; and *Availability* (of the physician), with the levels 3 days, 1 week, or 2 weeks (see Table 1). For experimental design and analysis, each attribute had 3–4 levels, and we assigned a value to each level of each attribute. For example, in the case of years of experience, we assigned two years a value of 1; four years a value of 2; six years a value of 3; and eight years has a value of 4. We assigned the same values to the rest of the levels for each attribute. With this numbering system, the stimuli set of physicians had a wide range: physicians with the worst total value (worst level of all characteristics, e.g., no specialty, long distance away, few years of experience) had a total of six points, whereas the physicians with the best values had a total of 11 points (see Table 1). We created 18 physicians with a value of 6, 18 physicians with a value of 11, and the rest of the total physician values

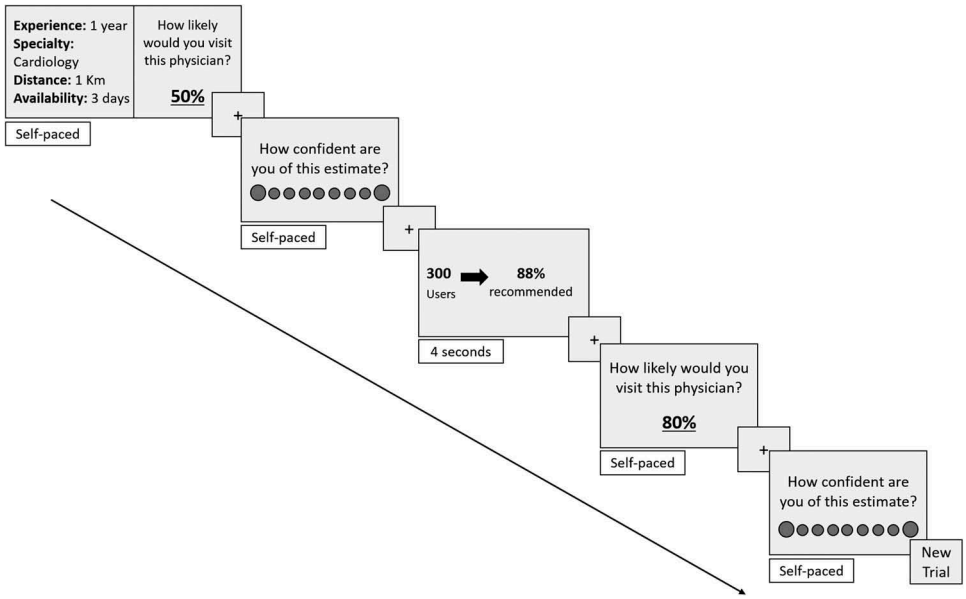

Figure 1. Participants saw six displays in each trial: (1) characteristics of the physician and first entry of likelihood to visit, (2) confidence rating entry, (3) display of recommendation, (4) second entry of likelihood to visit, (5) confidence rating entry, and (6) new trial. A fixation cross was shown between displays 1 and 4.

(7, 8, 9, and 10) had 21 physicians each. We balanced the conditions (low, high, and no recommendations) so that the same physicians were presented for each of the three conditions, but the trials were pseudo-randomized and physicians with the same value were not repeated immediately after each other. Importantly, participants did not see the values of the physicians, only the attributes. The values were only assigned to our fictitious physicians to systematize the trials, for example, to assure that participants experienced a broad array of physicians, and also to check that participants' first responses corresponded to the physician values. Physicians' gender was not presented to avoid possible biases.

The percentage of recommendations was also distributed equally in each condition. This means that the physicians corresponding to each value (i.e., 6–11) had recommendations that always resulted in an average of 50%. Moreover, the number of users was also balanced between conditions, which means that for each group of physician values, the low conditions always had an average of 10 users and the high condition had an average of 300 users (see Table 1).

We calculated different measures to analyze the results of our JAS paradigm such as the mean of first and second estimates, and the absolute differences between first and second estimates according to the three conditions. We used the absolute differences because our goal was to know how much the first estimate differed from the second not the relative direction of adjustment (up or down). Furthermore, we also calculated a Weight of Recommendation (WOR) index, which is based on the Weight of Advice index commonly found in the JAS literature (Bonaccio & Dalal, 2006). The WOR measures the influence of the recommendation on the participants' decisions, and it's created by the division of the "response difference" by the "opinion difference":

WOR = (second opinion – initial opinion)/(recommendation – initial opinion)

A WOR of one indicates that the participant adapted their estimate to exactly match the recommendation, while values of zero demonstrate that the estimate did not change. Negative values show that participants did not change their estimate toward the recommendation but went in the other direction. For instance, if participants indicate that they would visit a physician with a likelihood of 20%, and then they observe that only 30% of 300 users recommend that physician, the participants might change their second estimate to zero, meaning that they would not visit the physician at all. In this case, the WOR is −2.

Sociodemographic data

Before performing our task, participants also responded to sociodemographic questions that included age, gender, level of education, and occupation. To note, other tasks (Modified Card Sorting Test, Game of Dice Task) and a questionnaire (Big Five Inventory—short version) were also part of the experimental session, collected after the JAS paradigm, but the results of these were not included for the analysis of this paper.

Statistical analyses

Statistical analyses were performed using SPSS Statistics 23. An analysis of variance with repeated measures (ANOVA) was used to observe the relationship between the first response of likelihood to visit a physician and the physician's characteristics, and also for observing differences between conditions. Pairwise *t*-tests were used to compare the means of the different responses (e.g., first and second, after updating). In addition, a multivariate analysis of variance (MANOVA) was used to analyze potential gender effects.

Results

The descriptive values of all measures can be observed in Table 2. We used a repeated measures ANOVA to assess

Table 1. Physician characteristics, high and low conditions.

	High Condition				Low Condition			
Physician values	Number of raters	Average	Recommendation	Average	Number of raters	Average	Recommendation	Average
6	291	300	96%	50%	17	10	100%	50%
	311		85%		11		91%	
	280		55%		9		56%	
	290		45%		7		43%	
	337		15%		10		10%	
	291		6%		6		0%	
7	262	300	94%	50%	10	10	100%	50%
	294		83%		7		86%	
	265		68%		8		63%	
	292		51%		12		50%	
	306		34%		13		31%	
	345		14%		6		17%	
	336		8%		14		7%	
8	327	300	98%	50%	2	10	100%	50%
	309		81%		10		80%	
	308		71%		13		77%	
	316		49%		9		44%	
	257		28%		13		23%	
	269		18%		12		17%	
	314		6%		11		9%	
9	272	300	94%	50%	10	10	90%	50%
	289		82%		5		80%	
	275		69%		8		75%	
	302		51%		6		50%	
	354		32%		15		33%	
	326		18%		19		21%	
	282		2%		7		0%	
10	276	300	97%	50%	15	10	93%	50%
	296		81%		11		82%	
	321		62%		9		67%	
	347		46%		10		50%	
	306		36%		6		33%	
	294		20%		7		14%	
	260		6%		12		8%	
11	330	300	94%	50%	12	10	92%	50%
	298		85%		6		83%	
	282		54%		11		55%	
	286		45%		12		42%	
	307		20%		4		25%	
	297		4%		15		7%	

Physicians' values and their corresponding distribution among trials. The average of raters per value-group always results in 300 for the high condition, and in 10 for the low condition. The average of recommendations is always 50%.

Table 2. Descriptive results of the experimental paradigm.

	Range			
	Lowest	Highest	*M*	*SD*
Before advice rating	28.53	83.38	55.91	11.88
After advice rating	22.58	69.67	47.34	10.16
WOR—High	−.43	2.31	0.63	0.48
WOR—Low	−.85	1.59	0.45	0.39
Absolute difference—High	2.75	47.00	22.75	9.67
Absolute difference—Low	1.50	44.95	18.99	10.13
Absolute difference—Control	0.00	43.33	9.31	9.26

our first hypothesis, which aimed to reveal if the first response in our physician task (before seeing the recommendation of others) is related to the displayed characteristics of a physician (*Years of experience, Specialty, Availability*, and *Distance*—see Figure 2). Greenhouse-Geisser correction determined that the means of the estimates are statistically different among the value-groups (F (2.21, 286.00) = 412.88, $p < 0.001$). Post hoc Bonferroni correction also revealed significant pairwise differences between all value-groups with $p < 0.001$, except for the pair of values 6 and 7 ($p = 0.010$). The statistical means of the values 8 and 9 are also very similar (see Figure 2), nevertheless they also resulted in significant differences after Bonferroni correction ($p = 0.002$). With these results, we confirm that the first opinion of the users is in line with the provided information about the physicians and therefore we accept H1.

To assess our second hypothesis, we performed a pairwise *t*-test comparison between participants' estimates before and after the recommendations. Our analysis revealed a significant difference in the scores for the first (M = 55.91, SD = 11.88) and the second estimates (M = 47.34, SD = .89); t (129) = 12.59, $p < 0.001$, which indicates that participants were indeed influenced by the recommendations. Regarding our third hypothesis, repeated measures ANOVA of participants' estimates before and after receiving the recommendations revealed significant differences among the three conditions (low, high, control; see Figure 3); Greenhouser-Geisser corrections showed significant differences between the conditions (F (1.50, 194.07) = 200.20, $p < 0.001$). Post hoc Bonferroni corrections confirmed the differences at the

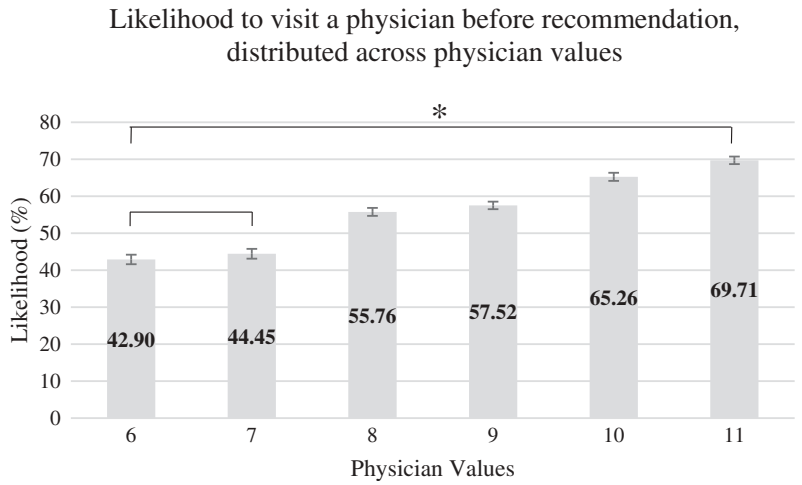

Figure 2. Likelihood to visit a physician before the recommendations were presented, distributed across the values assigned to physicians. All values were significantly different at pairwise level, after Bonferroni corrections ($p < 0.001$), except for the pairs of values 6 and 7 ($p = .010$). The mean differences are significant at the 0.05 level.

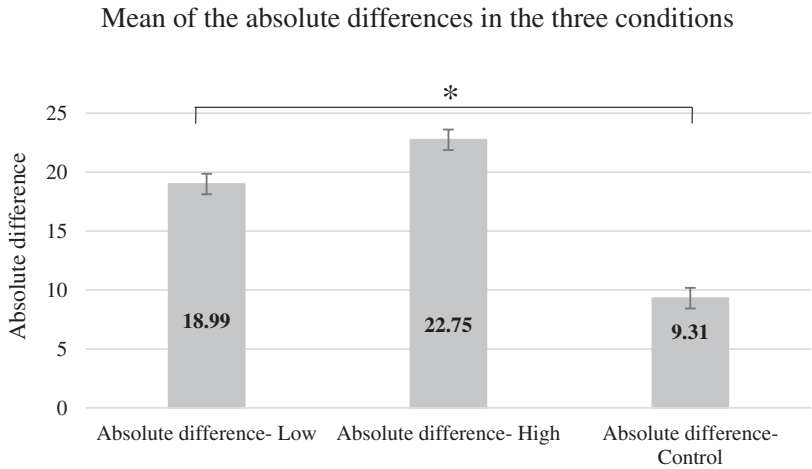

Figure 3. Mean of the absolute differences, of the experimental task in the three conditions. All values were significantly different at pairwise level ($p < 0.001$). control conditions.

pairwise level for all conditions ($p < 0.001$). Table 2 shows how the control condition ($M = 9.31$, $SD = 9.26$) differed the most in relation to the high ($M = 22.75$, $SD = 9.67$) and low ($M = 18.99$, $SD = 10.13$) conditions. These results demonstrate that participants were more influenced by a greater number of online reviewers. With this in mind, we accept Hypothesis 2 and 3.

Weight of recommendation

Confirmation of H3 shows that the difference between the first and second estimate was larger, when the number of raters was high. To investigate in which condition (low or high) the participants adapted their second estimate closer to the recommendation (H4), we used the WOR index. We found a significant difference when using pairwise comparison for the high ($M = .63$, $SD = .47$) and low ($M = .44$, $SD = .39$) conditions; $t(129) = 4.28$, $p < 0.001$. The mean of the high condition is closer to one, which indicates that the second estimate for the high condition, in comparison to the low condition, was closer to the recommendation. Therefore, we accept H4.

Confidence rating

To assess Hypothesis 5, which aimed to observe how confidence is related to the number of users rating, we performed the following analyses. First, we conducted a pairwise comparison. This analysis revealed no significant difference between the confidence ratings of the first ($M = 6.79$, $SD = 1.08$) and second estimates ($M = 6.83$ $SD = 1.18$); $t(129) = -.826$, $p = 0.41$. In a second step, we tested if there was a significant difference between participant confidence ratings across conditions, before and after the recommendation was presented. Our repeated measures ANOVA revealed a significant difference, after Greenhouser-Geisser corrections, between the three conditions for the confidence ratings of the first (F (1.33,

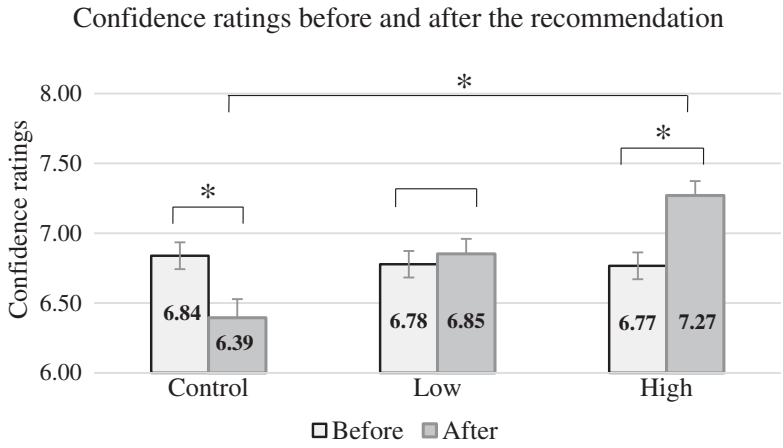

Figure 4. Confidence ratings before and after the recommendations. The confidence level increases when the number of users rating increases as well. For the second estimate, the confidence ratings in the three conditions were significantly different at pairwise level ($p < 0.001$).

171.95) = 52.25 $p < 0.001$) and second estimates (F (1.85, 238.18) = 63.92 $p = 0.003$—Figure 4). We found significant differences at the pairwise level for the second confidence ratings (High: $M = 7.27$, $SD = 1.18$; Low $M = 6.85$, $SD = 1.22$; Control: $M = 6.39$, $SD = 1.51$) in the three conditions ($p < 0.001$), after post hoc Bonferroni corrections. We therefore accept H5, as the confidence ratings for the second estimate were significantly higher when participants observed that the number of users rating a physician was high.

Further analysis

Although it was not part of our main research question, we analyzed potential gender effects in the sample. We compared the results of our dependent variables (JAS variables) with a MANOVA with gender as a fixed factor and found no significant differences, except for the absolute difference in the high condition, which was significantly higher in women ($M = 23.96$, $SD = 10.1$) than in men ($M = 20.05$, $SD = 8.07$) at a pairwise level ($p = 0.036$). No further significant differences were observed.

Discussion

The current study aimed to determine the influence of users' recommendations on physician rating websites. Overall, our results demonstrate that the number of recommendations that a physician has on his or her profile influences website users' decisions on whether to visit a physician or not, and this also changes users' confidence in their decisions.

With our first hypothesis, we tested whether the likelihood to visit a physician is related to the mere descriptions offered on the physician rating website. To achieve this, we designed the physician descriptions to vary in total, so they could be classified into six groups with different values according to their characteristics. Our first hypothesis was confirmed since we observed that participants' first response (likelihood to visit a physician) was significantly different among the six value-groups, with the mean likelihood to visit ascending with regard to the physicians' overall value. This means that before seeing the online recommendations, participants reported a higher likelihood to visit physicians with better characteristics. This result is important in two ways. First, the confirmation of our first hypothesis supports the validity of our experimental design, as we could confirm that a specific combination of physician's attributes can increase or decrease her/his perceived value. Second, we can also imply that in the absence of user-generated information, users guide their decisions accordingly to the mere description offered on the platform. In this sense, Carbonell and Brand (2017) showed that even though the specialty and the experience were the most important characteristics for assessing physician quality, most users base their decisions on subjective attributes of social media such as comments and ratings. Here lies the importance of understanding how others' physician ratings influence our decision to visit a physician: users of physician websites choose physicians based on the attributes shown to them, yet the reviews they see produce a significant change in their first "objective" decision, which is what we went on to address with our second hypothesis.

Building on the above, the second hypothesis tested if the recommendation of former patients causes a significant change in the likelihood to visit a physician. We observed a significant difference between participants' first and second responses, which demonstrates that user-generated recommendations on social media influence our decisions. In other words, one's first, "objective-like" impression can be significantly modified by the recommendation of others. This result is supported by previous studies which show that ratings and comments are important for these types of decisions (Carbonell & Brand, 2017; Flanagin & Metzger, 2013; Grabner-Kräuter & Waiguny, 2015), even though objective characteristics, such as specialty and experience might more-accurately assess physician quality (Carbonell & Brand, 2018). Thus, these results support the concerns of English physicians about the validity of these ratings (Patel et al., 2015). If users are able to estimate their likelihood to visit a physician based on his/her characteristics, but the ratings produce a significant change in their opinion, then physicians are right to be worried about the possible bias (Daskivich et al., 2018) that these ratings might induce. After all, marketing research has

also shown how some product ratings do not correspond to the product quality (De Langhe et al., 2016).

Our results demonstrate that decisions about which physician to visit can change drastically depending on the number of reviewers rating the physician on the internet. Our analysis with the WOR shows that a high number of reviewers influences participants in such a way that the second estimate is closer to the recommendation, when compared to the condition with fewer raters. This indicates that, users adapt their likelihood to visit a physician according to the recommendation and that this adaptation is more pronounced when the number of users rating the physician increases. Consequently, users of physician rating websites might be more likely to visit physicians who are positively rated by a large amount of users and less likely to visit badly rated physicians, as shown by Grabner-Kräuter and Waiguny (2015). Fortunately, different studies (Emmert et al., 2014; Kadry et al., 2011; Lagu et al., 2010) report that the majority of reviews and ratings in physician rating websites are positive. This does not mean that health-practitioners' worries are banal. On the contrary, our results allow us to infer that negative ratings can be very harmful for physicians. In this regard, the administrators of these platforms carry a major responsibility in controlling the veracity of the uploaded content. For instance, it is important to ensure that physician reviewers are indeed former patients; the comments should be respectful; and there must be a clear differentiation between the characteristics that influence the performance of the physician (e.g., specialty, experience, and time spent in the consult) and others that are not directly relevant (e.g., waiting times, friendliness of staff, and parking possibilities). With these measures, patients can make decisions with clear information that actually help them to discern which physician is more suitable for them.

Our analysis of participants' confidence ratings showed that when recommendations come from a high number of raters, people experience greater confidence in their judgment. With this in mind, we suggest that participants' confidence ratings can be interpreted as another measure of the credibility of the recommendation. This interpretation is in line with the studies of Flanagin and Metzger (2013), Grabner-Kräuter and Waiguny (2015), Lim and Van Der Heide (2015), which show that ratings from a high number of people are more credible than ratings from just a few. Based on our formulated hypotheses and the obtained results, we infer that the number of reviewers serves as a warranting cue, which not only influences the perceived credibility of the physician's profile, but the actual decision of which physician to visit.

Limitations

Although the results of this study are in line with previous research into physician rating websites and overall social influence, there are some caveats that need to be addressed. For example, we used the WOR in this experiment to indicate that users adapt their likelihood to visit a physician, making it closer to the recommendation when a high number of reviewers is presented. However, the use of this index has some limitations and issues to point out within the context that we have used it. First, the WOR computation from the JAS (normally termed the Weight of Advice index) is usually applied to a scenario where the estimate and the advice are exactly on the same scale, for instance, distance between two cities in km, price of a product, etc. In the current scenario, participants report the likelihood to visit a physician, but then see a percentage of recommenders. In this case, the advice is a recommendation, in percent, that participants get from other users who already visited the physician in question. Therefore, we are using two different scales, namely likelihood to visit a physician and percentage of recommendations. Moreover, we also limit the responses to a scale from 0 to 100%. Other JAS paradigms have no such limitation, which allows participants to be more flexible in their answers. The scenario that we used limited the response because of ecological validity; percentages below 0 and above 100 do not exist on physician rating websites. Finally, the use of the WOR computation is usually limited to positive values and we did not do this. In other versions of the JAS, when computing the Weight of Advice, it is odd when participants go in the opposite direction of the advice (if they do not agree with the advice, they should stay with their first opinion but not go away from the advice because they have not been given any information that would suggest they should move their opinion in the opposite direction). In our paradigm, however, it is plausible that users go in the opposite direction of the recommendation, yet this still speaks for the influence of the recommendation. For instance, if a participant indicates that they would visit a physician with a likelihood of 20%, and then they observe that only 30% of 300 users recommend that physician, the participant might change their second estimate to be lower than 20%. In this case, the user actually moves their estimate away from the recommendation and we calculate a negative WOR. This negative WOR would be dismissed in another JAS study; however, in our paradigm, it makes sense to use our index in this way, because of the two scales that we previously mentioned, and therefore, we included trials with a negative WOR.

Conclusions and future research

From a cognitive perspective, many investigations have shown that judges trust the advice provided by experts more than advice provided by novices (Meshi et al., 2012; Önkal et al., 2017; Reyt et al., 2016; Swol & Sniezek, 2005). Similarly, communication research has shown that a large number of reviews on a product or service have a similar "expertise" effect (source credibility) on the opinion of the users (Flanagin & Metzger, 2013; Lim & Van Der Heide, 2015). The current study allows us to infer that the cognitive process underlying advice taking might be similar to the one we go through when seeing recommendations on the Internet, as our JAS allowed us to assess the influence of this warranting cue. Future research may examine this parallel further, contributing to our understanding of the impact that users' reviews have on our decision making. For instance, our paradigm could be performed in an fMRI scanner. This would reveal if the neural mechanism for integrating expert advice during a decision, as found by Meshi et al. (2012), is similar to receiving a recommendation from many users *versus* fewer

users. We expect that, as observed in the current study, the effect of a higher number of users is similar to the effect of advice coming from an expert. In another line, researchers could investigate the motivation for reviewing and sharing opinions online. For example, Meshi et al. (2016) examined the regions of the brain involved in the sharing self-related information online (Meshi et al., 2016), and this could be directly relevant for understanding users reviewing physicians on these platforms. Indeed, capitalizing on neuroimaging methods, to better understand the interactions offered by social media, has already been stressed previously (Meshi, Tamir, & Heekeren, 2015). Even though the scientific community is advancing in this new direction, more investigation needs to be done in order to understand cognition in the era of social media.

In brief, physician rating websites are commonly used by physicians, patients and social media users looking for a physician. Even though their importance might be increasing, there are concerns regarding the validity of the ratings provided by former patients, since these might not reflect the actual quality of a physician. This study showed that users are able to objectively assess which physicians are "better" by reporting that they would visit them according to their characteristics. Importantly, their opinion changed significantly after seeing the ratings of others. This change was more pronounced when the number of users rating the physician was higher. The concern of physicians about the validity of these ratings is therefore valid and needs to be addressed by these platforms. Moreover, these platforms are useful tools for patients and users searching for a physician. We showed how the number of ratings have influence on this decision making process and it is necessary to keep investigating other factors that contribute to the understanding of this process, which ultimately will benefit the patients.

Although some studies show that physician ratings are generally positive, we have shown that negative ratings could have a direct negative influence in the decision to visit a physician. In this regard, the administrators of these platforms should pay special attention to the content that users upload.

Funding

This work was supported by the German Research Foundation (DFG) under grant No. [GRK 2167], Research Training Group "User-Centred Social Media."

References

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 127–151. doi:10.1016/j.obhdp.2006.07.001

Brehmer, B., & Hagafors, R. (1986). Use of experts in complex decision making: A paradigm for the study of staff work. *Organizational Behavior and Human Decision Processes*, *38*, 181–195. doi:10.1016/0749-5978(86)90015-4

Brooks, A. W., Gino, F., & Schweitzer, M. E. (2015). Smart people ask for (my) advice: Seeking advice boosts perceptions of competence. *Management Science*, *61*, 1421–1435. doi:10.1287/mnsc.2014.2054

Carbonell, G., & Brand, M. (2018). Choosing a physician on social media: Comments and ratings of users are more important than the qualification of a physician. *International Journal of Human–Computer Interaction*, *34*, 117–128. doi:10.1080/10447318.2017.1330803

Daskivich, T. J., Houman, J., Fuller, G., Black, J. T., Kim, H. L., & Spiegel, B. (2018). Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. *Journal of the American Medical Informatics Association*, *25*, 401–407. doi:10.1093/jamia/ocx083

De Langhe, B., Fernbach, P., & Lichtenstein, D. (2016). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research*, *42*(6), 817–833. doi:10.1093/jcr/ucv047

DeAndrea, D. C., Van Der Heide, B., Vendemia, M. A., & Vang, M. H. (2015). How people evaluate online reviews. *Communication Research*, 1–18. doi:10.1177/0093650215573862

Emmert, M., Meier, F., Heider, A. K., Dürr, C., & Sander, U. (2014). What do patients say about their physicians? An analysis of 3000 narrative comments posted on a German physician rating website. *Health Policy*, *118*, 66–73. doi:10.1016/j.healthpol.2014.04.015

Emmert, M., Sauter, L., Jablonski, L., Sander, U., & Taheri-Zadeh, F. (2017). Do physicians respond to web-based patient ratings? An analysis of physicians' responses to more than one million web-based ratings over a six-year period. *Journal of Medical Internet Research*, *19*, e275. doi:10.2196/jmir.7538

Flanagin, A. J., & Metzger, M. J. (2013). Trusting expert-versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior*, *29*, 1626–1634. doi:10.1016/j.chb.2013.02.001

Gao, G., McCullough, J. S., Agarwal, R., & Jha, A. K. (2012). A changing landscape of physician quality reporting: Analysis of patients' online ratings of their physicians over a 5-year period. *Journal of Medical Internet Research*, *14*, e38. doi:10.2196/jmir.2003

Grabner-Kräuter, S., & Waiguny, M. K. (2015). Insights into the impact of online physician reviews on patients' decision making: Randomized experiment. *Journal of Medical Internet Research*, *17*, e93. doi:10.2196/jmir.3991

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*, 117–133. doi:10.1006/obhd.1997.2697

Kadry, B., Chu, L. F., Kadry, B., Gammas, D., & Macario, A. (2011). Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating. *Journal of Medical Internet Research*, *13*, e95. doi:10.2196/jmir.1960

Lagu, T., Hannon, N. S., Rothberg, M. B., & Lindenauer, P. K. (2010). Patients' evaluations of health care providers in the era of social networking: An analysis of physician-rating websites. *Journal of General Internal Medicine*, *25*, 942–946. doi:10.1007/s11606-010-1383-0

Lim, Y.S, & Van Der Heide, B. (2015). Evaluating the wisdom of strangers: the perceived credibility of online consumer reviews on yelp. *Journal Of Computer-mediated Communication*, *20*(1), 67-82.

Lin, X., Spence, P. R., & Lachlan, K. A. (2016). Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior*, *63*, 264–271. doi:10.1016/j.chb.2016.05.002

Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How expert advice influences decision making. *PLoS One*, *7*, e49748. doi:10.1371/journal.pone.0049748

Meshi, D., Mamerow, L., Kirilina, E., Morawetz, C., Margulies, D. S., & Heekeren, H. R. (2016). Sharing self-related information is associated with intrinsic functional connectivity of cortical midline brain regions. *Scientific Reports*, *6*. doi:10.1038/srep22491

Meshi, D., Tamir, D. I., & Heekeren, H. R. (2015). The emerging neuroscience of social media. *Trends in Cognitive Sciences*, *19*, 771–782. doi:10.1016/j.tics.2015.09.004

Önkal, D., Gönül, M. S., Goodwin, P., Thomson, M., & Öz, E. (2017). Evaluating expert advice in forecasting: Users' reactions to presumed vs. experienced credibility. *International Journal of Forecasting*, *33*, 280–297. doi:10.1016/j.ijforecast.2015.12.009

Patel, S., Cain, R., Neailey, K., & Hooberman, L. (2015). General practitioners' concerns about online patient feedback: Findings from a descriptive exploratory qualitative study in England. *Journal of Medical Internet Research*, *17*, e276. doi:10.2196/jmir.4989

Reyt, J. N., Wiesenfeld, B. M., & Trope, Y. (2016). Big picture is better: The social implications of construal level for advice taking. *Organizational Behavior and Human Decision Processes*, *135*, 22–31. doi:10.1016/j.obhdp.2016.05.004

Rui, J. R. (2017). Source–target relationship and information specificity: Applying warranting theory to online information credibility assessment and impression formation. *Social Science Computer Review*, *36*, 331–348. doi:10.1177/0894439317717196

Sniezek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, *17*, 173–190. doi:10.1002/bdm.468

Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology*, *44*, 443–461. doi:10.1348/014466604X17092

Tong, S. T., Van Der Heide, B., Langwell, L., & Walther, J. B. (2008). Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook. *Journal of Computer-Mediated Communication*, *13*, 531–549. doi:10.1111/j.1083-6101.2008.00409.x

Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in. In M. Knapp & J. Daly (Eds.), *Handbook of interpersonal communication* (Vol. 3, pp. 529–563). Thousand Oaks, CA: SAGE.

Walther, J. B., Liang, Y. J., Ganster, T., Wohn, D. Y., & Emington, J. (2012). Online reviews, helpfulness ratings, and consumer attitudes: An extension of congruity theory to multiple sources in Web 2.0. *Journal of Computer-Mediated Communication*, *18*, 97–112. doi:10.1111/j.1083-6101.2012.01595.x

Walther, J. B., Van Der Heide, B., Hamel, L. M., & Shulman, H. C. (2009). Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using Facebook. *Communication Research*, *36*, 229–253. doi:10.1177/0093650208330251