Movement and Nutrition in Health and Disease

Assessing the quality of weight loss information on the German language web

| Research

Selina Meyer, David Elsweiler, Bernd Ludwig
Chair for Information Science, University of Regensburg, 93040 Regensburg, Germany
Correspondence: selina.meyer@ur.de; david.elsweiler@ur.de; bernd.ludwig@ur.de

Received 27 March 2020; Accepted 30 March 2020; Published 31 March 2020

Abstract: This article examines the quality of weight loss information on the German language web and studies how websites, likely to be accessed via popular web search engines, are evaluated by end users. Sixty-five websites were identified and qualitatively examined with respect to content quality as defined by the literature, as well as meta information on design and structure of the page. In a further step, the same web pages were evaluated by non-expert users in an online study. Deficiencies were found, both in terms of the quality of information on the websites, and with respect to the search behaviour and the rating competence of users. Many of the examined web pages showed little or no relevance for weight loss and 46% of the pages covered a maximum of only 3 of the 18 content criteria. Significant differences in results were identified for websites of different type. Media websites covered most criteria (M = 5.5, SD = 2.66), followed by commercial sites (M = 4.10, SD = 2.54). Nonprofit sites contained the fewest content criteria (M = 2.72, SD = 2.7), but made the least number of unsubstantiated claims and met the most design criteria. In the majority of cases, agreement between participant ratings was found to be poor to moderate. They also generally found fewer content criteria than the gold standard suggested, but gave higher quality ratings and underestimated the proportion of unsubstantiated claims. We conclude that users have low expectations for weight loss information on the Internet or are influenced by criteria other than content when assessing quality.

Keywords: Weight-loss; quality of information; web search; credibility assessment; information seeking.

1. Introduction

Obesity rates in Germany have grown rapidly over the past decades. While only 37% of the German population were overweight in 2005 [1], by 2015, 67.1% of adult men and 53% of adult women were overweight [2]. The number of obese individuals has almost doubled in the same time period. Being overweight, in particular when this is to the level of obesity, is currently one of the largest health risks, as it increases the risk of diabetes, cancer and cardiovascular diseases, making this a huge societal problem [1].

The state endorses measures that inform on the advantages of a healthy and active lifestyle in order to prevent obesity [2]; the Internet being a primary channel for distributing such information. Ninety percent of Germans use the Internet [3] and two thirds of Internet users older than ten years search for health information online [4]. As such, the Internet is a corner stone of health education for German adolescents and adults [5].

The Internet, however, also poses great risks as misinformation can be distributed at an alarming rate [6]. This is also true of weight loss and nutrition information as previous investigations have evidenced [7–9]. Most Google searches mainly return commercial or media sites as opposed to official or solely informational sources, but as previous research has shown, commercial sites are often qualitatively inferior to others [10]. This discrepancy between website types may also apply to weight loss information, since commercial sites pursue different goals to non-profit or media sites. There is a large amount of weight loss information available on the web and the incorrect application of such information can have detrimental consequences. Inadequate diets for weight loss can lead to long term weight gain or even cause eating disorders [1,11,12]. Information overload and the large quantities of low quality information on the Internet pose a problem, particularly for health related topics, since people do not look for such information on a regular basis and therefore have neither specific trusted sources, nor are they proficient in properly evaluating a sites' quality [8,13].

Previous examinations of health and weight loss information on the Internet have shown that there is a lack of quality and consistency in content [14–17]. Most of the current literature, however, either evaluated only structural criteria or so-called "credibility indicators" [16] or focused on a narrow subject within weight loss information such as bariatric surgery [17]. In their work, Modave and colleagues [15] showed that only one fifth of English language websites retrieved with topical queries on the Google search engine contained more than 50% of relevant information pertaining to nutrition, physical activity and behaviour changes and found vast differences in the amount of available information on different weight loss related topics. Unfortunately, most people rely on the search engine's ranking when looking for information [18, 19]. Given the low quality of weight loss information on the Internet, this approach is precarious. Furthermore the health literacy of half of the German population can be classified as problematic [9], resulting in difficulties when judging the quality of health information. Relevant literature has found large differences between expert and non-expert credibility judgements of American health sites, with non-expert users judging more subjectively than experts [20]. Thus, it is important to assess the quality of accessible weight loss information on the Internet to detect potential deficiencies and examine how users perceive the quality of such information.

Although attaining information in their native language is one of the most important factors for Internet users in Germany [21], there are no current studies dealing with the quality of weight loss information on German websites. This study examines the quality of weight loss information on German websites using a mixed-methods approach. The goal being to answer the following research questions (RQ):

RQ1: What is the quality of accessible weight loss information on German websites?

RQ2: Are certain weight loss topics covered more adequately than others?

RQ3: Is there a difference in quality across website types?

RQ4: How do users evaluate websites pertaining to weight loss?

RQ5: Which criteria influence users in their evaluation?

2. Methods

To achieve a representative pool of weight loss queries and websites likely to be submitted and received we applied a multi-staged process. An initial study identified common search engine queries pertaining to weight loss. These were submitted to Google Trends, which provided the most popular related queries submitted to Google in the previous year. These queries were then submitted to Google to identify high ranking websites returned.

Sixty-five retrieved websites were subsequently examined qualitatively with respect to content quality, as well as meta information such as design and page structure. The same pages were then evaluated by non-expert users, in order to assess the average user's rating competency.

2.1. Identifying common search queries and results

Relevant queries were generated using two simulated work task situations [22] outlining information needs pertaining to weight loss in an initial online study:

Situation 1:

At your last appointment with your General Practitioner, he pointed out that your BMI is in the overweight range. He listed various health risks associated with greatly increased weight and advised you to adopt a healthier lifestyle to lose weight. After the appointment, you want to look for different ways to lose weight on the Internet.

Situation 2:

You have tried multiple weight loss diets, but always struggled with the yoyo effect. You find it difficult to integrate weight loss diets in your everyday life, as they are usually associated with increased effort (i.e. counting calories). After the last diet failed again, you decided to change your approach. Instead of a short term weight loss, that will not last, you want to attempt a lifestyle change, that will help you lose weight in the long term. To do this, you first want to find information on how to live healthier and lose weight in the process on the Internet.

The recruitment process was tailored such that individuals who could relate to the scenarios took part. Two sources of participants were a self-help group for obese people and a group preparing for bariatric surgery. We also included individuals who were not obese, since we expected anyone to be eligible to look for weight loss information online, as long as they are not satisfied with their current weight, regardless of their actual body mass index. Twenty-six of the 40 participants were female. Participants were aged between 20 and 71 years old (M = 42, SD = 12.29). 72.5% of participants described themselves as currently wanting to lose weight. Sixty-five percent had attempted to lose weight within the past six months. The 147 resulting queries were coded and assigned to thematic categories. Using Google Trends (https://trends.google.de/trends/?geo=DE), we identified the 26 most commonly searched queries in Germany over the past year for each of these categories, such that the proportion with which the categories occurred in the initial study was kept (see Table 1). We used Google Trends as this was successfully applied in previous studies of websites on the English and Spanish language web [15, 23].

The resulting queries were submitted to Google, the most used search engine in Germany [24]. For each query the first three retrieved sites were considered for evaluation. Since sponsored sites are less frequently opened than others [25], we only included non-sponsored search results. A few sites had to be discarded as they were non-informational in nature or only locally relevant. One site appeared in three of the results pages. The 67 resulting sites were separated by provider and classified into three categories: commercial (20), media (29) and non-profit (18). Ten of the 18 non-profit sites were Wikipedia articles. Two of 67 sites (one media, one commercial) were taken offline over the course of the study and could, therefore, not be included in the quality evaluations in part 4.

Table 1. Chosen queries by category and category frequency

Category	No. Codes (No. mentions)	Most frequent queries	Approximate English translation
Vegetarian/Vegan	6 (7)	vegetarisch	vegetarian
Physical activity	7 (7)	Muskelaufbau	muscle building
Obesity and medicine	9 (13)	Schilddrüse Insulin	thyroid insulin
BMI	4 (5)	BMI	BMI
Diet	12 (16)	Diät Diätplan	diet diet plan
Dietary change	7 (16)	Ernährungsumstellung Ernährung umstellen	diet change change diet
Nutrition	31 (45)	Fasten gesund essen gesunde Ernährung Kalorientabelle Rezepte zum Abnehmen	fasting eat healthy healthy nutrition calorie table weight loss recipes
Specific diets and techniques	13 (18)	Low Carb Intervallfasten Keto	low carb interval fasting keto
Yoyo effect	6 (10)	Jojo-Effekt	yoyo effect
Weight loss generally	28 (53)	Abnehmen Übergewicht schnell abnehmen abnehmen ohne Sport Gewichtsreduktion gesund abnehmen	losing weight overweight lose weight fast lose weight without sports weight reduction lose weight healthily
Health	4 (6)	gesund leben	live healthily
Keep weight	2 (4)	Gewicht halten	keep weight

2.2. Evaluation criteria

The criteria, against which pages were evaluated were derived from current weight loss guidelines by the WHO [26, 27], the German DGE [28] and DAG [29], as well as a related study [15] serving as content criteria pertaining to *nutrition*, *physical activity*, *behavioural changes*, *pharmacotherapy* and *surgical options*. The structure was adapted from the evaluation form of a recent study on the quality of English language websites relating to dieting and weight loss [15]. Content criteria were included in form of checkboxes. Criteria were, for example, balancing the energy input and output (*nutrition*), strengthening major muscle groups twice a week (*physical activity*) or self-monitoring (*behaviour change*). The complete evaluation form can be found in the appendix.

A five-point Likert scale indicated the quality of each topic, ranging from "no mention" over "poor", "average" and "good" to "very good". We also included certain design criteria, such as the usage of hyperlinks, the structure of the site and the colour scheme and the share of unsubstantiated claims in each category and overall in the evaluation form. The quality rating for each topic depended on the number of criteria covered by the site and the share of unsubstantiated claims regarding the topic.

All websites were evaluated against these criteria twice by the lead author. The two evaluations were conducted two months apart and tested for intra-rater reliability to ensure the validity of the evaluations. Checkboxes were tested using Cohen's Kappa, Likert scales were tested with squared Cohen's Kappa. Mean agreement was 0.76 (*SD* = 0.17) on checkboxes and 0.61 (*SD* = 0.34) on Likert scales. 31 sites scored almost perfect results of 0.8 or higher on checkboxes. Agreement on Likert scales was only in the almost perfect range for 20 sites, however, the median Kappa score on Likert scales was 0.72, indicating that agreement on half of the sites was rather high. In a further iteration, individual points were reconsidered by rereading the corresponding articles, in order to achieve perfect agreements and end up with a single evaluation per site. These evaluations were used as gold standard for user evaluations in later analyses.

The evaluation form was then distributed to non-expert users by means of an online survey. Participants were asked to answer a few personal questions regarding their health literacy and behaviour before rating websites. Each participant was allocated one article from each category at random and in a random order. Participants were asked to rank the three viewed sites by preference at the end of the study. They also had the chance to describe what positively or negatively influenced their quality ratings for each site and why they ranked the sites in a certain way. The online study was shared to SurveyCircle (https://www.surveycircle.com/de/) and a Facebook group for people with obesity. The survey was online between August 17, 2019 and September 30, 2019.

2.3. Meta data

Beyond the content criteria, relevant meta data were also recorded for all websites. This included information regarding the author of the site and their qualification, the publication date, the length and readability and the presence of advertisements. **Authorship:** 55% of sites (eight commercial, 14 media, 14 non-profit) gave no information on the author. In 26% of cases (eight commercial, ten media), the author was named, but there was no mention of their qualification. Nevertheless, all but seven sites (three commercial, four media) provided contact information. **Date:** 21 articles were published or updated in 2019, eight in 2018, eight were older. The remaining sites provided no publication date. **Length:** Articles were between 536 and 13,420 words in length. Non-profit sites were longest on average (*M* = 3079.667, *SD* = 3070), followed by commercial sites (*M* = 2776.687, *SD* = 1370.4). Media sites were shortest (*M* = 2506, *SD* = 1537.42). Due to the large standard deviation, a Kruskall-Wallis test did not show significant differences in length between website types. **Readability:** Readability was determined using an online text analysis tool (http://www.schreiblabor.com/textanalyse/) to calculate the Flesch Reading Ease (Flesch Index) [30] adapted for German texts. Results varied between 12 (hard to read) and 69 (easy to read). Non-profit sites were hardest to read (*M* = 34.11, *SD* = 8.1), while media (*M* = 51.83, *SD* = 11.25) and commercial sites (*M* = 47.4, *SD* = 8.17) were easier. A one-way ANOVA showed a significant main effect $F(2) = 19.24$, $p < .001$, $\eta_p^2 = .375$ across website types, with significant differences between non-profit and other sites ($p < .001$). Since there is an expectation that non-profit sites are higher quality than other types, this is problematic, especially considering the low health literacy in Germany [9]. Wikipedia articles yielded the lowest readability scores (*M* = 32.44, *SD* = 6.42). **Advertisement:** It was documented whether the sites and the main text were ad-free. Since content-relevant adverts can negatively impact a sites' perceived credibility [31], we also determined, whether adverts were relevant to the sites' content. We also determined if any of the sites could be interpreted as so-called advertorials, editorial articles, with the intent to advertise a certain product. These types of articles are seldom recognised as commercial content by readers and are as such more successful than traditional advertising [32]. Sixteen sites (ten commercial, six media) were identified as advertorials, and only 23 sites were completely ad-free (two commercial, four media, 17 non-profit). Of the 44 sites with ads, 24 had advertisements in the main text and 18 had at least partially targeted ads. Ninety percent of commercial and 86% of media sites included some form of advertisement.

3. Results

Generally, the results pertaining to quality content were disappointing. No single site covered more than 55.56% of content criteria. Only in three cases a quality rating of "very good" concerning single content subjects was given. *Nutrition* and *behavioural change* were especially prone to unsubstantiated claims.

3.1. Website evaluation

Out of the five high-level content criteria themes, *nutrition* was the best covered subject. Still, on average only two of four criteria were detected (*SD* = 1.15). Only 25% of sites covered more than two recommended items. 47 sites recommended to focus on specific foods, 45 to avoid specific foods, and 32 covered the balance of energy input and output. Only six of the evaluated sites gave the advise to limit salt intake.

Fifty percent of sites covered one of the items recommended for *behavioural change*, 25% covered three or more of the five items. The most commonly covered item was improving diet or physical activity (36 sites), followed by finding and addressing barriers to change (24 sites). Eighteen sites discussed behavioural management activities and strategies to maintain lifestyle changes, and 16 discussed self-monitoring. On average, 1.68 (SD = 1.65) single criteria items were mentioned.

Only 25% of sites gave at least one recommendation pertaining to *physical activity*, with 11 sites recommending the strengthening of major muscle groups two times a week. Less than one item was discussed on average (*M* = 0.54, *SD* = 0.77).

Only two sites gave information about *pharmacotherapy* and three about *surgical options*. Media sites generally covered the most criteria regarding *nutrition*, *physical activity* and *behavioural change*. It was also apparent that the retrieved non-profit sites, in particular, were occasionally non-relevant to the subject and often covered no criteria at all (see Table 2).

Table 2. Average number of covered criteria per topic and site type: mean (SD)

	Nutrition	Physical activity	Behaviour change
Commercial	2.00 (1.25)	0.42 (0.84)	1.68 (1.67)
Media	2.39 (0.87)	0.71 (0.81)	2.36 (1.70)
Non-profit	2.03 (1.15)	0.39 (0.61)	0.61 (0.85)

A Kruskall-Wallis test was employed to determine differences in the overall number of criteria covered between site types. There was a significant main effect $H(2) = 10.79$, $p = .004$, $\eta_p^2 = .142$, which was further investigated using a Dunn posthoc test. Differences were significant between non-profit and media sites (*p* = .003), with media sites covering the most subjects (*M* = 5.5, *SD* = 2.66) and non-profit sites the fewest (*M* = 2.72, *SD* = 2.7). Commercial sites (*M* = 4.1, *SD* = 2.54) showed no significant differences to other site types. Across all types, only 4.32 (*SD* = 2.84) out of 18 criteria were covered on average. Only two sites managed to cover 10 criteria (one commercial, one media). 46.15% covered less than four.

3.1.1. Unsubstantiated claims

Unsubstantiated claims were most common among media sites with regard to single topics. Only four non-profit sites contained unsubstantiated claims. On the other hand, only three sites of media and commercial, respectively, did not contain any unsubstantiated claims.

Topics that were covered more extensively were also more likely to include unsubstantiated claims. Commercial ($r_\tau = .438$, $p < .015$) and media sites ($r_\tau = .530$, $p < .001$) both had a strong correlation between the number of criteria covered and the number of unsubstantiated claims across subjects. The overall percentage of unsubstantiated claims was fairly high across all types of sites, with non-profit sites having the smallest share of unsubstantiated claims on average (all sites: *M* = 63.37, *SD* = 38.66, media: *M* = 70.54, *SD* = 36.88, commercial: *M* = 78.89, *SD* = 28.81, non-profit: *M* = 35.83, *SD* = 37.82). Often, no sources were indicated at all, causing all claims to be judged as unsubstantiated.

While a lot of these claims were rather harmless in nature, some can be deemed as manipulative, often causing unrealistic expectations. As an example, one site blamed dairy products for illnesses, such as respiratory infects and chronic headaches [33], another introduced a diet that would supposedly cause a weight loss of five kilograms within the first week [34], with none of the sources cited proving more than three kilograms of weight loss per week [35,36]. This could cause a rapid loss of motivation if the promised results are not achieved, leading to people blaming themselves for their "failure" to lose weight as expected.

3.1.2. Quality ratings

As in Modave et al. [15], the quality ratings were derived from the number of covered criteria and the unsubstantiated claims. Quality was thus rated highest for the topics *nutrition* and *behaviour change*. Only in three cases, however, was a rating of "very good" achieved, meaning that all criteria of a topic were covered. A rating of "very good" was given once on the topic *nutrition* (commercial), and twice on *behaviour change* (commercial, media). In 28 cases, the rating was "good". None of the sites achieved good ratings across all topics. Most sites were rated as "average" across topics, with the exception of *pharmacotherapy* and *surgical options*, which were rarely covered at all (see Figure 3). Only one site was rated "average" on *pharmacotherapy* (non-profit) and *surgical options* (media), respectively. Two sites received a rating of "poor" for *pharmacotherapy* (one media, one non-profit) and one for *surgical options* (non-profit).

3.1.3 Design

Non-profit sites fulfilled most of the seven design criteria (M = 4.44, SD = 0.51), followed by commercial (M = 3.26, SD = 1.28) and media sites (M = 3.11, SD = 1.03). The Kruskal-Wallis test yielded a significant main effect $H(2) = 20.249$, $p < .001$, $\eta_p^2 = .294$. The posthoc Dunn test showed significant differences between non-profit sites and the other types ($p < .001$). Non-profit sites were most likely to include proper citation (61.1%) and hyperlinks (61.1%), have a competent author (44.4%) and use relevant and adequate graphics (44.4%). Media sites most often had minimal page layering (46.4%). Commercial sites did not perform best on any of the design criteria. 90.8% of sites had a distinguishable structure and 98.4% had appropriate font and background colours, without major differences between site types.

3.2. Survey results

In total, 103 people took part in the online survey, 95 through the platform SurveyCircle. Due to the differing sizes of the website type groups, commercial and non-profit sites were rated between six and seven, media between three and four times. Participants were aged between 16 and 90 years old (M = 29.27, SD = 11.48). 73.7% of participants were female. The majority were students (62.1%) and had some kind of college degree (64.2%). 27.2% were employed. The high percentage of female participants aligns with relevant literature, which shows that women are more interested in weight loss than men [37]. 72.8% of participants were younger than 30 years and as such belong to the group of heavy Internet users in Germany [38]. Only 11 participants claimed that they never use the Internet to look for weight loss information. The majority of participants indicated, that they had trouble losing or keeping their weight and almost half (49 participants) admitted to trying trend diets in the past. Participants were asked 17 questions pertaining to their health behaviour, information behaviour, previous knowledge on the subject, and weight satisfaction.

On average, participants took 17.5 minutes to complete the survey. There was a weak correlation, between the overall length of allocated texts and time spent on the survey ($r_\tau = .095$, $p = .013$), however, the length of individual articles had no impact on the time participants spent rating them. Similarly, the readability of texts did not influence the time spent on the rating.

3.2.1. Agreement

To determine inter-rater reliability, the evaluations were divided into different sets. The purpose behind this was to find out on what level of granularity participant agreement was highest.

Set 1 included only the single criteria in form of checkboxes, and had the potential to show whether participants agreed on the presence or absence of the individual criteria. Since many of the criteria were similar to others and open for interpretation to a certain degree, agreement on this set was expected to be rather low. Agreement on set 1 was measured using Fleiss' Kappa.

Set 2 was made up of sub scores pertaining to the different topics. It consisted of the number of checked criteria per topic, the total score of quality ratings, the quality ratings for the entire site, and the percentage of unsubstantiated claims for the entire site. Since distinguishing between individual criteria may have been a hard task for participants, we assessed whether agreement on the number of checked criteria per topic was higher. This would show that participants agreed on the presence or absence of certain information, even if they struggled to reliably assign labels to information into individual criteria. The same applies to unsubstantiated claims. It is cognitively challenging to classify claims as belonging to a certain topic, which is why we only included the percentage assigned for the entire site in this set.

Quality ratings were expected to be more subjective than the other parts of the survey. We therefore included the total score of quality ratings given for the topics and the quality ratings given by users on the entire site.

Set 3 consisted of all total scores (total number of checked content criteria, number of checked design criteria, total score of quality ratings). With this set we aimed to discover to what extent users agreed on the overall content and quality of sites, independent of content categories and unsubstantiated claims. Set 2 and set 3 were evaluated using Krippendorff's Alpha for ratio scales and multiple raters. Another factor was the time spent on the website, and how it influences the inter-rater reliability.

Generally, Set 1 yielded the lowest and Set 2 the highest inter-rater reliability scores. This demonstrates that users, although often unsure about single criteria, do agree about how well a site covers different topics overall. Further tests showed that agreement on unsubstantiated claims and quality scores on single criteria was especially poor. The average inter-rater reliability score never reached more than 0.61 on either quality or unsubstantiated claims and dropped as low as -0.13 for unsubstantiated claims.

A possible explanation for this is that properly determining the percentage of unsubstantiated claims (particularly in longer web sites) would take more time than most participants were willing to give, thus participants simply provided subjective estimates.

Agreement was lowest on non-profit sites, on all sets but Set 1, including agreement on unsubstantiated claims and quality only, regardless of the time participants spent rating the sites. A possible reason for this might be the hard readability of the non-profit sites compared to others.

Removing participants who spent less time on each site before providing their scores resulted in higher agreement. Doing this systematically, as depicted in Figure 1, results in an initial peak around the 100 second mark. It is likely, that participants who spent less than 100 seconds per site, were not paying sufficient attention to properly complete the survey. As a result, they were excluded in all further analyses. At this mark, inter-rater reliability scores for Set 2 ranged between .199 and .917, with a mean of .59 (*SD* = .18), which is slightly lower than the cut-off point to substantial agreement (.61).

To find out whether agreement was influenced by participants' health behavior or previous knowledge, we tested for correlations between agreement and the answers given by participants at the beginning of the survey. None of the correlations were significant. However, a high readability score positively correlated with higher agreement (r_s = .39, *p* < .001). Moreover, ad-free sites had lower agreement (r_s = -.33, *p* < .001). Five participants also achieved high agreement ratings (> .8) on two sites, while seven participants had especially low ratings (< .375) twice. This indicates that health behavior and previous knowledge has at least some effect on the reliability of the evaluation.

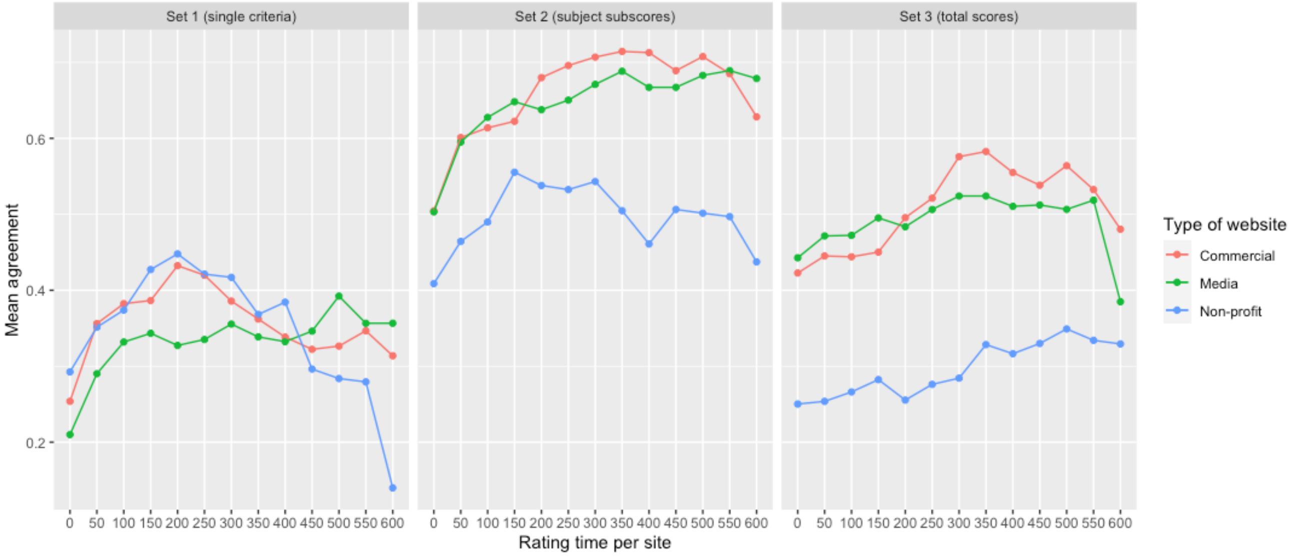


Figure 1. Agreement on different sets depending on the time spent rating sites

3.2.2. Relevance, credibility and ranks

Twenty-one sites were deemed non-relevant for weight loss by all participants (five commercial, seven media, nine non-profit). Media sites were seen as relevant in 41.78% of cases, commercial sites in 34.25% and non-profit sites in 21.33%. 23 sites were not seen as credible by all participants (14 media, seven commercial, two non-profit). Pearson's Chi-squared test showed significant differences between judgements and site type for both relevance ($X^2(2, N = 227) = 7.45, p = .024$) and credibility ($X^2(2, N = 227) = 13.74, p = .001$).

There were significant differences in relevance between media and non-profit sites ($p = .027$) and credibility differences between non-profit and commercial ($p = .013$), as well as non-profit and media sites ($p < .001$). In 52% of cases, non-profit sites conveyed new knowledge to individual participants, while only 37.97% of media and 32.88% of commercial sites managed to do the same. Overall, it became apparent that the majority of the sites retrieved in the study are not able to provide new knowledge, even when a user's health literacy is comparatively low, as the sites often only convey surface-level knowledge or are irrelevant to the actual information need.

When ranking the evaluated websites according to preference, participants ranked non-profit sites highest on average ($M = 1.91, SD = 0.84$), followed by media sites ($M = 1.97, SD = 0.79$). Commercial sites were ranked lowest ($M = 2.12, SD = 0.82$). Even though there was no significant main effect, it is notable that non-profit sites were ranked highest, even though they were the most frequent to be judged irrelevant.

3.2.3. Reasons for quality judgements and rankings

Participants were given the chance to explain why they had a positive or negative impression of a site in free-text fields. Their answers were coded by content using the R package RQDA (http://rqda.r-forge.r-project.org). Each code was then allocated to a category. The resulting categories in descending order of frequency: Content, quality, sources, design, and type of website. The most common codes and their frequencies are outlined in Table 3. Participants criticised, for example, a lack of relevant sources, a lack of relevant content, a lack of in-depth information, a lack of critical reflection, or a lack of respectability. Some participants commented on specific citations given by the articles, evaluating their quality and reputability. Sites that offered new knowledge, practical tips or a good overview of the topic, used good graphics and were well structured were praised. Some participants were also guided in their judgement by their previous opinion on certain sites, as for example Wikipedia.

Table 3. Most common reasons for participants' impressions of a site

Code	Mentions	Category
Missing sources	32	Sources
Valuable information	24	Content
Missing reputability	21	Quality
Proper indication of sources	21	Sources
Appearance	19	Design
Irrelevant	19	Content
Missing/superficial information	19	Content
Little benefit	19	Content
Reputability	17	Quality
Advertisement/sale	16	Quality
No new knowledge	15	Content
Website/provider	14	Site type
Confusing structure	14	Design
Good structure	13	Design
Readability/interest	13	Quality
Gives overview	12	Content
Diverse information	12	Content
One-sided information	11	Content
Comprehensive	9	Content

3.3. Gold standard vs participant evaluation

After examining the participants' evaluations and the inter-reliability between participants, participants' ratings were compared to the gold standard in order to find specific areas, where Internet users misjudge online content.

Including the gold-standard evaluations caused a slight increase of inter-rater reliability across all three sets and the quality ratings. On the other hand, agreement dropped when only the unsubstantiated claims were tested (see Table 4). A Kruskall-Wallis test showed, that none of the differences in agreement were significant. Nevertheless, differences can be observed between the rating behaviour of participants and the author.

Table 4. Inter-rater reliability scores between participants with and without including the gold standard: mean (SD)

	Set 1	Set 2	Set 3	Quality	Unsubstantiated claims
Commercial					
Participants only	0.38 (0.15)	0.61 (0.18)	0.44 (0.19)	0.35 (0.34)	0.16 (0.22)
incl. gold standard	**0.44 (0.13)**	**0.63 (0.18)**	**0.47 (0.23)**	**0.38 (0.30)**	**0.03 (0.12)**
Media					
Participants only	0.33 (0.18)	0.63 (0.17)	0.47 (0.22)	0.31 (0.40)	0.13 (0.30)
incl. gold standard	**0.42 (0.14)**	**0.65 (0.14)**	**0.49 (0.18)**	**0.47 (0.27)**	**-0.01 (0.21)**
Non-profit					
Participants only	0.37 (0.23)	0.49 (0.18)	0.27 (0.20)	0.16 (0.31)	-0.07 (0.13)
incl. gold standard	**0.44 (0.19)**	**0.50 (0.15)**	**0.27 (0.20)**	**0.26 (0.34)**	**-0.12 (0.10)**

Participants tended to select fewer checkboxes (M = 3.27, SD = 2.38) than the author (M = 4.32, SD = 2.84), but the totals of the quality scores assigned by participants were higher (participants: M = 7.4, SD = 4.13, author: M = 3.81, SD = 2.57) on average. Participants also underestimated the share of unsubstantiated claims (participants: M = 43.85, SD = 31.15, author: M = 65.75, SD = 38.24). The Kruskall-Wallis test showed significant main effects between participants and the author for checkboxes ($H(1) = 7.54$, $p = .006$, $\eta_p^2 = .018$), as well as quality ratings ($H(1) = 43.168$, $p < .001$, $\eta_p^2 = .113$) and unsubstantiated claims ($H(1) = 18.168$, $p < .001$, $\eta_p^2 = .046$). Among the checkboxes, the strongest differences were visible in the topic *nutrition* among non-profit sites and the topics *nutrition*, *physical activity* and *behavioural change* among the media sites (see Figure 2).

Quality ratings differed mostly with respect to *nutrition*, *physical activity* and *behaviour change* across all site types (see Figure 3). *Pharmacotherapy* and *surgical options* showed fewer differences between raters, as they were not mentioned at all on most sites.

Differences in evaluation of unsubstantiated claims between the author and participants were especially high for commercial and media sites (see Figure 4). The findings presented in this section paint a picture of superficial participant rating behaviour leading to disagreement with the author's more rigorous analysis.

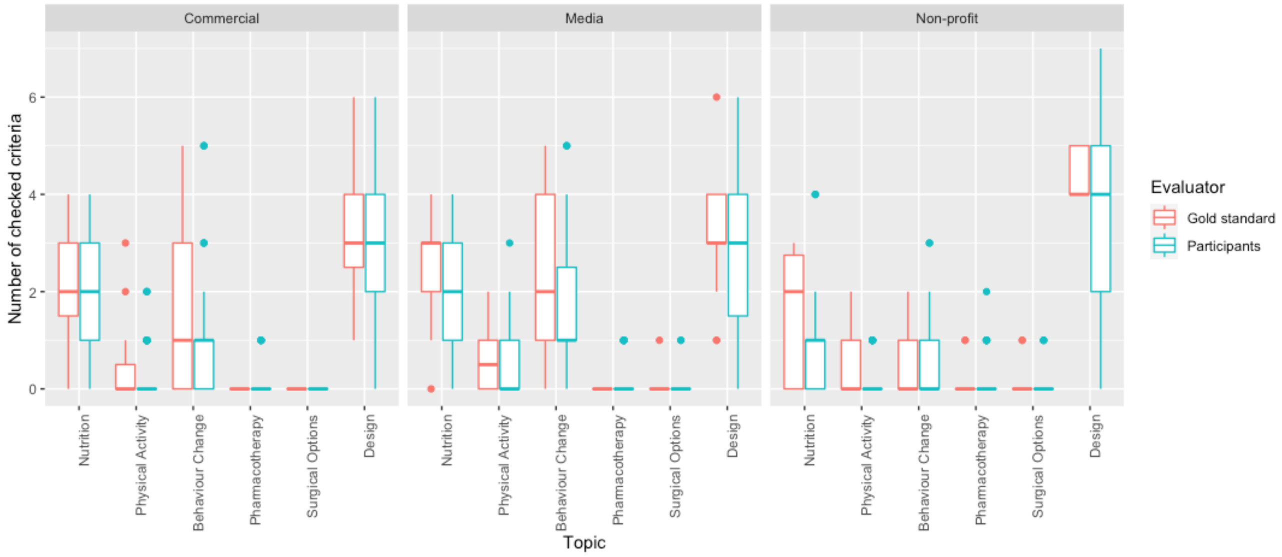


Figure 2. Number of Checkboxes checked by participants and in gold standard across website types

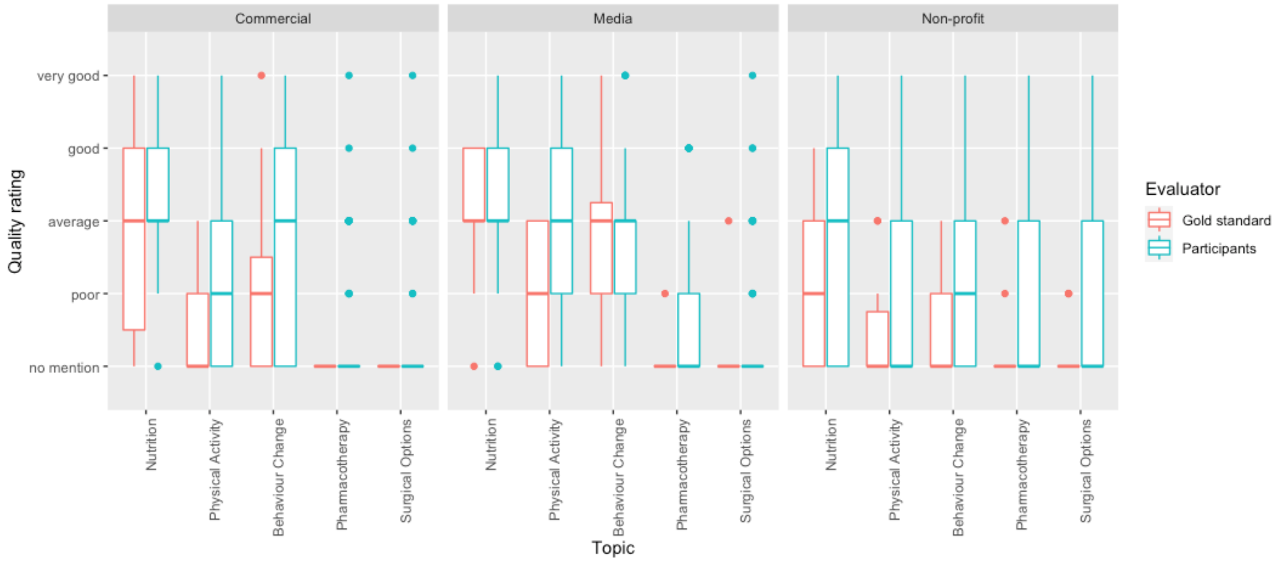


Figure 3. Quality ratings given by participants and in gold standard across website types

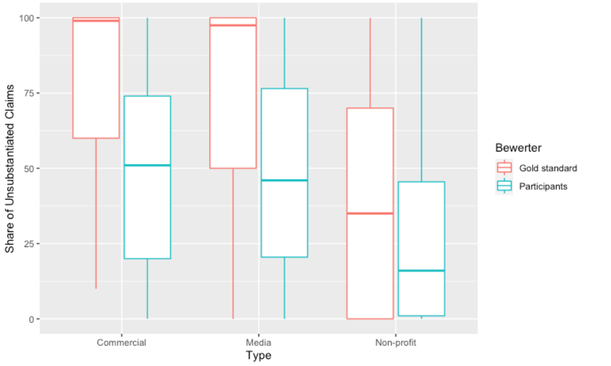


Figure 4. Share of unsubstantiated claims indicated by participants and the gold standard across sites

4. Discussion

This section reflects on the presented results in the context of our research questions.

RQ1: What is the quality of accessible weight loss information on German websites?

The sample of 65 retrieved websites, likely to be found when looking for weight loss information on Google demonstrates that the quality of such information in the German language web is far from perfect. Some of the sites covered a few of the criteria recommended for weight loss by official sources, however, they often propagated nonscientific information and contain a large share of unsubstantiated claims. None of the sites covered more than 55.56% of the recommendations and only few of the sites achieved a quality rating higher than "average". This aligns with similar findings for English and Spanish language websites [15,23].

Another outcome is that many of the sites returned by Google were not relevant to the topic at all. This also reflects on some of the queries collected in the initial study (i.e. "thyroid"), which are more focused on medical reasons for overweight rather than a healthy diet or lifestyle changes. Since most of the queries were also rather short, consisting of only one or two words, they did not give enough context for a search engine to return relevant results. Simply changing the query "thyroid" to "thyroid overweight" might return more relevant results. This reveals another possible difficulty in providing and finding weight loss information online, namely a lack of Internet users' information retrieval skills.

RQ2: Are certain weight loss topics covered more adequately than others?

While the previous study for English language websites indicated that mainly nutrition and physical activity are covered in weight loss [15], German websites seem to place additional emphasis on behavioural change. That is changes which can offer users the chance to achieve a long-term lifestyle change. Information on pharmacotherapy and surgical options was scarce, even though these topics would be especially valuable to people with severe obesity. In addition, only few sites provided sufficient information about physical activity, an integral part to a healthy lifestyle. Many sites recommended to do "enough" sports, however few elaborated on what "enough" actually means. One explanation for this could be the high rate of female participants in the initial study which sourced the seed queries, as women are more likely to change their nutrition in order to lose weight, while men are more likely to change their activity levels [39].

RQ3: Is there a difference in quality across website types?
Even though many of the non-profit sites retrieved were irrelevant to weight loss, these provided appropriate sources for their claims most often, had the fewest advertisements and were deemed most credible by participants. On the other hand, much of the information they offered was not really suitable for daily life and as such they did not seem very useful to participants, and were more suited as a general overview. Commercial sites covered more criteria but provided many unsubstantiated claims. Media sites covered the most criteria overall. The results reveal that it is difficult to find relevant, German non-profit sites using popular queries and the challenging readability of non-profit sites may make it even harder to comprehend the information provided. Moreover, just four of the sites returned were published by government-funded organisations.

These findings underline the importance of working towards making more official resources available through common google queries, so users need not rely on commercial sites, which are mainly driven by sales, and other non-scientific information.

RQ4: How do users evaluate websites pertaining to weight loss?
There were major differences between the gold standard and participants' ratings, as well as across the ratings provided by participants themselves, which again aligns with existing literature indicating a difference in quality judgements between users and experts [20], and that users have difficulties evaluating health related websites [8,40].

Participants particularly disagreed on single checkbox items and quality ratings, although they often selected a similar number of checkboxes for each topic. The impression won is that users generally agree on whether a site offers information on a topic or not, but struggle with assigning information to specific checkbox items or making the same quality judgements on that basis.

Since previous studies showed a link between health literacy and evaluation [9,39], we expected the health behaviour to have a certain impact on agreement, however there were no significant correlations found in that regard. Nevertheless, the fact that participants sometimes had an especially high or an especially low inter-rater reliability with the same individuals more than once, suggests that individual differences do play some role in users' proficiency when judging online content pertaining to weight loss.

Participants selected fewer checkboxes regarding content criteria across almost all site types and topics than the expert rater, but still rated webpages as being of higher quality. They also underestimated the share of unsubstantiated claims. This suggests, that participants were satisfied more easily even with low quality information.

RQ5: Which criteria influence users in their evaluation?
Participants' free-text responses show their judgements were most often based on content criteria and sources. Some participants reported looking at specific sources and judged their quality and reputability or criticized sites, when no sources were referenced. The literature shows, that people often report that these are the primary criteria they apply when judging a website [18,41], but rarely actually pay attention to sources or fact check in practice [18]. This is also mirrored in participants' quality ratings, which which judged pages to be of relatively high quality compared to the gold standard, even when a large percentage of the information on a page was unsubstantiated. This indicates that the relationship between a lack of sources and the quality of the content was not apparent to participants in most cases and that the share of unsubstantiated claims did not influence their quality ratings.

A further finding was that participants valued articles, which offered new information and specific strategies for weight loss that can be incorporated into daily life. Similar results can be found in the literature, which indicated that users mainly look for comprehensive information and day to day strategies online [20,40]. Similarly, the role of subjective quality judgements and participants' perception of the validity of the information given on the sites played a role and has been previously found [20]. The comparatively high quality ratings indicate that these criteria were not fully used in practice, even though participants' were aware of a lack in quality.

5. Limitations

There are some limitations to our study that should be acknowledged. The queries from the simulated work task situations were evaluated using Google Trends. Despite Google Trends being a valuable tool that offers insight from a naturalistic population that has been used in equivalent investigations in the past (e.g. reference 15), the tool offers no contextual information. This may have resulted in some queries, not relevant to diet and weight-loss to be included in the study. We refer again to the example of the "thyroid" query from the initial study which, google trends confirmed was a frequently submitted query to Google over the past year. It is likely, however, that many of these searches were unrelated to weight loss. We know from the initial study that this query can be plausibly submitted for weight loss information, but since only 40 participants were studied it is not possible to generalise this to the general population. While this is a limitation in terms of the generalisability of the web pages analysed, it does highlight a lack of search competencies, which search engines must deal with. If users search using these queries, not all of the web pages returned will be relevant to their tasks.

6. Conclusion

The Internet will continue to increase in importance as a source of health information, which is problematic considering the low quality of information we found in this study. A lack of relevance of many sites and the low number of government-funded sites, which would have the most reliable information, are especially concerning. On average, only 24% of criteria were covered and no site covered more than ten out of 18 possible criteria. Future work can build on these findings and the corpus of evaluated websites in order to facilitate the access of relevant and high quality weight loss information on the Internet in the long term. Since all sites were judged both by Internet users, and the lead author who has in-depth knowledge of literature on obesity and weight loss recommendations, we were able to unveil strengths and weaknesses in the ability of users to judge a website's quality. In future work these results will be used to discover whether it is possible to predict user judgements for perceived credibility and relevance based on a site's meta criteria and content and study how different user groups judge these websites differently.

Conflict of interest

The authors declare no conflict of interest.

References

1 Pudel V, Ellrott T. Adipositas – ein gesellschaftspolitisches Problem? Chirurg 2005; 76: 639–646.

2 Robert-Koch-Institut. Gesundheit in Deutschland, 2015. DOI: 10.1055/s-2007-993182.

3 Statistisches Bundesamt (Destatis). 90 % der Bevölkerung in Deutschland sind online, 2018. Available: https://www.destatis.de/DE/Presse/Pressemitteilungen/2018/09/PD18_330_634.html. Accessed March 27, 2020.

4 Anonymous. Zwei von drei Internetnutzern suchen nach Gesundheitsinformationen. Deutsch Ärztebl 2016. Available: https://www.aerzteblatt.de/nachrichten/66227/Zwei-von-drei-Internetnutzern-suchen-nach-Gesundheitsinformationen. Accessed March 27, 2020.

5 Hirschfelder G. Wege aus der Digitalisierungsfalle – Ernährungskommunikation und Ernährungsbildung. In: Ernährung im Fokus. Bonn: Bundeszentrum für Ernährung, 2018 (09-10): 284–288. Available: https://www.bzfe.de/_data/files/5885_2018_eif_Leseprobe.pdf. Accessed March 27, 2020.

6 Lazer DMJ, Baum MA, Benkler Y, et al. The science of fake news. Science 2018; 359: 1094–1096.

7 Cusack L, Desha LN, Del Mar CB, Hoffmann TC. A qualitative study exploring high school students' understanding of, and attitudes towards, health information and claims. Health Expect 2017; 20: 1163–1171.

8 Eysenbach G. Credibility of health information and digital media: new perspectives and implications for youth. In: Metzger MJ, Flanagin AJ, eds. Digital media, youth, and credibility. Cambridge MA: MIT Press, 2008: 123–154.

9 World Health Organization - Regional Office for Europe. Health literacy: The solid facts, 2013. Available: http://www.euro.who.int/pubrequest. Accessed March 27, 2020.

10 Kunst H, Khan KS. Quality of web-based medical information on stable COPD: comparison of non-commercial and commercial websites. Health Info Libr J 2002; 19: 42–48.

11 French SA, Story M, Downes B, Resnick MD, Blum RW. Frequent dieting among adolescents: psychosocial and health behavior Ccorrelates. Am J Public Health 1995; 85: 695–701.

12 Wimmer-Puchinger B. Adipositas und Essstörungen im Brennpunkt - Eine Auseinandersetzung mit dem Einfluss von Wirtschaft und Gesellschaft auf Kinder und Jugendliche. Wien: Wiener Programm für Frauengesundheit, 2015. Available: https://www.wien.gv.at/gesundheit/beratung-vorsorge/frauen/frauengesundheit/pdf/adipositas-essstoerungen.pdf. Accessed March 27, 2020.

13 Eysenbach G, Jadad AR. Evidence-based patient choice and consumer health informatics in the internet age. J Med Internet Res 2001, 3:, e19.

14 Central Versicherung. Praxis Dr. Internet - Studie zum Krankheitssuchverhalten in Deutschland sowie zur Qualität von Gesundheitsinformationen im Internet, 2015. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjwjp6l2dPpAhUEQxUIHSAQBOYQFjAAegQIARAB&url=https%3A%2F%2Fwww.central.de%2Fresource%2Fblob%2F33860%2F5956a59a7f151952

ee2d6d10547d033d%2Fergebnisbericht-data.pdf&usg=AOvVaw3L8lqL-ViGSRMgkayJ_hjV. Accessed March 27, 2020.

15 Modave F, Shokar NK, Peñaranda E, Nguyen N. Analysis of the accuracy of weight loss information search engine results on the internet. Am J Public Health 2014; 104: 1971–1978.

16 Guardiola-Wanden-Berghe R, Gil-Pérez JD, Sanz-Valero J, Wanden-Berghe C. Evaluating the quality of websites relating to diet and eating disorders. Health Info Libr J 2011; 28: 294–301.

17 Vetter D, Ruhwinkel H, Raptis DA, Bueter M. Quality assessment of information on bariatric surgery websites. Obes Surg 2018; 28: 1240–1247.

18 Eysenbach G, Köhler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002; 324: 573–577.

19 Pan B, Hembrooke H, Joachims T, Lorigo L, Gay G, Granka L. In Google we trust: users' decisions on rank, position, and relevance. J Comput-Mediat Comm 2007; 12: 801–823.

20 Stanford J, Tauber E, Fogg BJ, Marable L. Experts vs. online consumers: a comparative credibility study of health and finance web sites, 2002. Available: http://www.ebusinessforum.gr/old/content/downloads/comparativeCredibilityStudy.pdf. Accessed March 27, 2020.

21 Birkmann C, Dumitru RC, Prokosch HU. Evaluation of health-related internet use in Germany. Methods Inf Med 2006; 45: 367–376.

22 Borlund P. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Inf Res 2003; 8. Available: http://informationr.net/ir/8-3/paper152. Accessed March 27, 2020.

23 Cardel MI, Chavez S, Bian J, Peñaranda E, Miller DR, Huo T, Modave F. Accuracy of weight loss information in Spanish search engine results on the internet. Obesity (Silver Spring) 2016; 24: 2422–2434.

24 Rabe L. Statistiken zu Suchmaschinen, November 2019. Available: https://de.statista.com/themen/111/suchmaschinen/. Accessed March 27, 2020.

25 Jansen BJ, Resnick M. An examination of searcher's perceptions of nonsponsored and sponsored links during ecommerce Web searching. J Am Soc Inf Sci Technol 2006; 57: 1949–1961.

26 World Health Organization. Physical activity, February 2018. Available: https://www.who.int/news-room/fact-sheets/detail/physical-activity. Accessed: March 27, 2020.

27 World Health Organization. Healthy diet, October 2018. Available: https://www.who.int/news-room/fact-sheets/detail/healthy-diet. Accessed: March 27, 2020.

28 Deutsche Gesellschaft für Ernährung. 10 Regeln der DGE. Available: https://www.dge.de/fileadmin/public/doc/fm/10-Regeln-der-DGE.pdf. Accessed March 27, 2020.

29 H. Hauner H, Moss A, Berg A, et al. Interdisziplinäre Leitlinie der Qualität S3 zur „Prävention und Therapie der Adipositas". Adipositas - Ursachen, Folgeerkrankungen, Therapie 2014; 8: 179–221.

30 Jacobsen J, Zitzelsberger A. Kann man die Benutzerfreundlichkeit von Text messen? In: Bosenick T, Hassenzahl M, Müller-Prove M, Peissner M, eds. Usability Professionals. Stuttgart: Fraunhofer Verlag, 2006: 66–69.

31 Zha W, Wu HD. The impact of online disruptive ads on users' comprehension, evaluation of site credibility, and sentiment of intrusiveness. Am Commun J 2014; 16: 15–28.

32 Wojdynski BW. The deceptiveness of sponsored news articles: how readers recognize and perceive native advertising. Am Behav Scientist 2016; 60: 1475–1491.

33 Rehberg C. Gesunde Ernährung, gesund essen, March 2020. Available: https://www.zentrum-der-gesundheit.de/gesunde-ernaehrung-die-regeln.html. Accessed March 27, 2020.

34 Widhammer-Zintl J. Schnell abnehmen laut Wissenschaft: 3 simple Regeln, die wirklich helfen. Available: https://www.instyle.de/beauty/schnell-abnehmen. Accessed March 27, 2020.

35 Brehm BJ, Seeley RJ, Daniels SR, D'Alessio DA. A randomized trial comparing a very low carbohydrate diet and a calorie-restricted low fat diet on body weight and cardiovascular risk factors in healthy women. J Clin Endocrinol Metab 2003; 88: 1617–1623.

36 Johnstone AM, Horgan GW, Murison SD, Bremner DM, Lobley GE. Effects of a high-protein ketogenic diet on hunger, appetite, and weight loss in obese men feeding ad libitum. Am J Clin Nutr 2008; 87: 44–55.

37 Institut für Demoskopie Allensbach. Fast jeder zweite Deutsche würde gerne abnehmen. Allensbacher Kurzbericht, 10. April 2014. Available: https://www.ifd-allensbach.de/fileadmin/kurzberichte_dokumentationen/PD_2014_08.pdf. Accessed March 27, 2020.

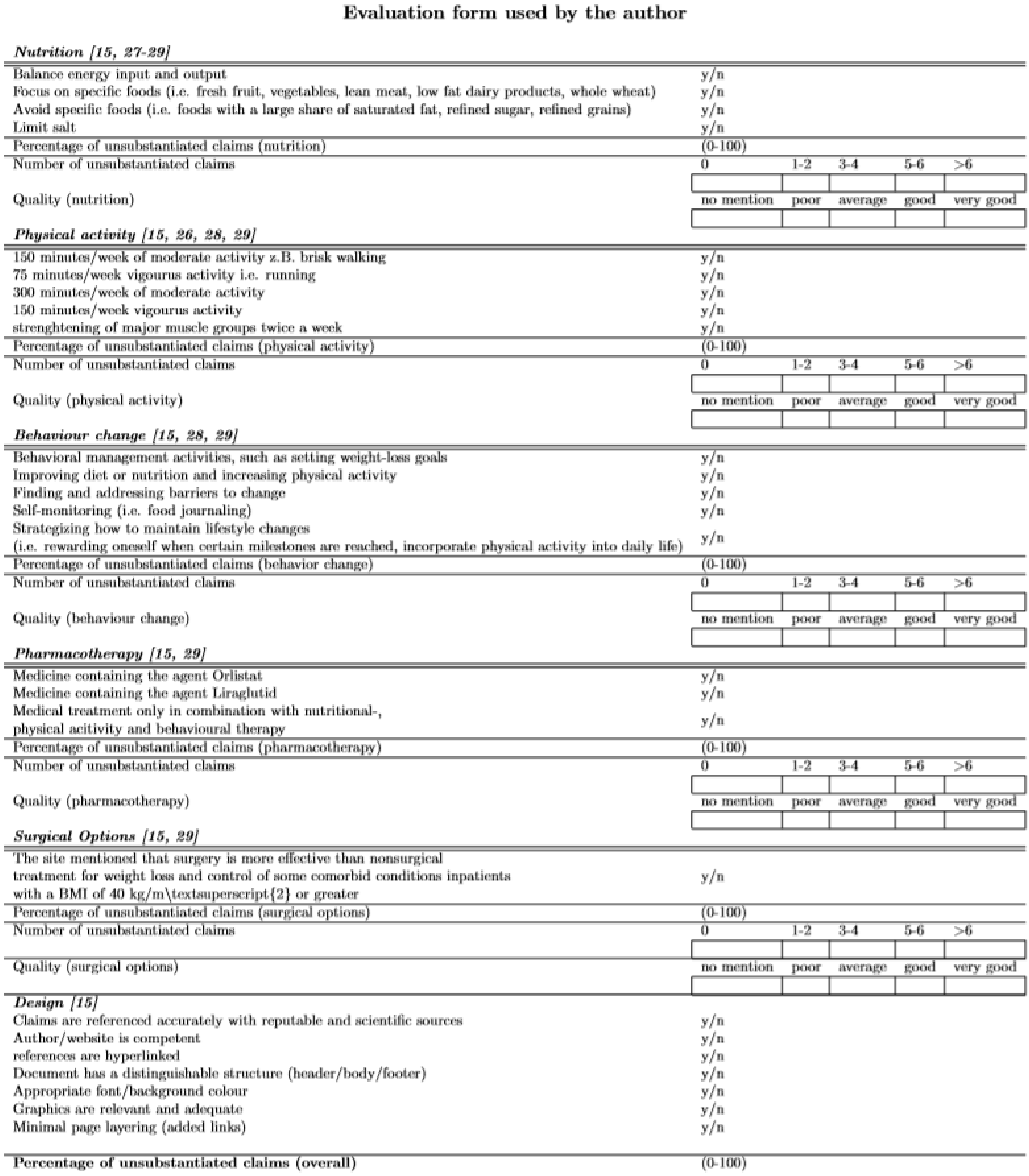
38 Projektgruppe ARD/ZDF-Multimedia. ARD/ZDF Onlinestudie 2917 – Kern-Ergebnisse. Available: http://www.ard-zdf-onlinestudie.de/files/2017/Artikel/Kern-Ergebnisse_ARDZDF-Onlinestudie_2017.pdf. Accessed March 27, 2020.

39 Jordan S, Hoebel J. Gesundheitskompetenz von Erwachsenen in Deutschland – Ergebnisse der Studie „Gesundheit in Deutschland aktuell" (GEDA). Bundesgesundheitsbl 2015; 58: 942–950.

40 Holmberg C, Berg C, Dahlgren J, Lissner L, Chaplin JE. Health literacy in a complex digital media landscape: pediatric obesity patients' experiences with online weight, food, and health information. Health Informatics J 2019; 25: 1343–1357.

41 Vervier L, Calero Valdez A, Ziefle M. Should I trust or should I go?" or what makes health-related websites appear trustworthy? An empirical approach of perceived credibility of digital health information and the impact of user diversity. In: Proceedings of the 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health, 2018: 169–177. DOI: 10.5220/0006734401690177.

Appendix



Evaluation form used by the author

Nutrition [15, 27-29]

Balance energy input and output	y/n				
Focus on specific foods (i.e. fresh fruit, vegetables, lean meat, low fat dairy products, whole wheat)	y/n				
Avoid specific foods (i.e. foods with a large share of saturated fat, refined sugar, refined grains)	y/n				
Limit salt	y/n				
Percentage of unsubstantiated claims (nutrition)	(0-100)				
Number of unsubstantiated claims	0	1-2	3-4	5-6	>6
	□	□	□	□	□
Quality (nutrition)	no mention	poor	average	good	very good
	□	□	□	□	□

Physical activity [15, 26, 28, 29]

150 minutes/week of moderate activity z.B. brisk walking	y/n				
75 minutes/week vigourus activity i.e. running	y/n				
300 minutes/week of moderate activity	y/n				
150 minutes/week vigourus activity	y/n				
strenghtening of major muscle groups twice a week	y/n				
Percentage of unsubstantiated claims (physical activity)	(0-100)				
Number of unsubstantiated claims	0	1-2	3-4	5-6	>6
	□	□	□	□	□
Quality (physical activity)	no mention	poor	average	good	very good
	□	□	□	□	□

Behaviour change [15, 28, 29]

Behavioral management activities, such as setting weight-loss goals	y/n				
Improving diet or nutrition and increasing physical activity	y/n				
Finding and addressing barriers to change	y/n				
Self-monitoring (i.e. food journaling)	y/n				
Strategizing how to maintain lifestyle changes (i.e. rewarding oneself when certain milestones are reached, incorporate physical activity into daily life)	y/n				
Percentage of unsubstantiated claims (behavior change)	(0-100)				
Number of unsubstantiated claims	0	1-2	3-4	5-6	>6
	□	□	□	□	□
Quality (behaviour change)	no mention	poor	average	good	very good
	□	□	□	□	□

Pharmacotherapy [15, 29]

Medicine containing the agent Orlistat	y/n				
Medicine containing the agent Liraglutid	y/n				
Medical treatment only in combination with nutritional-, physical acitivity and behavioural therapy	y/n				
Percentage of unsubstantiated claims (pharmacotherapy)	(0-100)				
Number of unsubstantiated claims	0	1-2	3-4	5-6	>6
	□	□	□	□	□
Quality (pharmacotherapy)	no mention	poor	average	good	very good
	□	□	□	□	□

Surgical Options [15, 29]

The site mentioned that surgery is more effective than nonsurgical treatment for weight loss and control of some comorbid conditions inpatients with a BMI of 40 kg/m^{2} or greater	y/n				
Percentage of unsubstantiated claims (surgical options)	(0-100)				
Number of unsubstantiated claims	0	1-2	3-4	5-6	>6
	□	□	□	□	□
Quality (surgical options)	no mention	poor	average	good	very good
	□	□	□	□	□

Design [15]

Claims are referenced accurately with reputable and scientific sources	y/n
Author/website is competent	y/n
references are hyperlinked	y/n
Document has a distinguishable structure (header/body/footer)	y/n
Appropriate font/background colour	y/n
Graphics are relevant and adequate	y/n
Minimal page layering (added links)	y/n
Percentage of unsubstantiated claims (overall)	(0-100)