Comparing the effectiveness of different fact-check formats

Using the truth sandwich and explicit falsehood labels

Tulin, M.; Hameleers, M.; de Vreese, C.

Publication date
2022

Document Version
Final published version

Research | 20.10.2022

Comparing the Effectiveness of Different Fact-check Formats

Using the Truth Sandwich and Explicit Falsehood Labels

Introduction

Mis- and disinformation are widely considered to be threatening our democracies. At the same time, a growing number of individual studies and meta-analyses have shown that fact-checks can help to correct misperceptions and false beliefs [1]. The effectiveness of fact-checks has mainly been measured as the extent to which exposure to corrective messages lowers misperceptions in experimental settings [2]. The question remains how we should conceptualize the effectiveness of fact-checks: What is the desired impact of corrective messages, and which formats are best equipped to establish these impacts?

These questions form the backbone of our latest BENEDMO experiment, which is the first of a series of experiments in the digital research lab of BENEDMO. The aim of this lab is to comprehensively map the effectiveness of fact-checking and provide evidence-based recommendations for interventions against disinformation. In the first experiment, we explored the effectiveness of different fact-checking formats. Together with fact-checking experts from the

BENEDMO network, we constructed fact-checks that varied in the location and prominence of false information. The fact-checks either repeated the false claim and then debunked it or they wrapped it up in truthful information, which is also referred to as the truth sandwich [3]. Second, some fact-checks explicitly mentioned labels that communicate a clear verdict on the falsehoods of verified statements, while others left it up to the reader to draw their own conclusions. Third, we looked beyond the direct effects on belief correction by exploring the extent to which exposure to fact-checks can stimulate critical thinking, and whether exposure to corrective information also has an effect on detecting subsequent falsehoods and the selection of fact-checks of other dubious claims.Together, our first BENEDMO experiment aims to comprehensively explore the impact of the salience of references to falsehoods in fact-checking information and their effects on critical behaviors that help people to navigate other information in their newsfeed.

Practically, it can be argued that corrective information is successful when its impact can transgress the evaluation of specific debunked falsehoods. Fact-checks take substantial resources to conduct, and the time and resources invested to fact-check one false statement greatly exceeds the time it takes to disseminate falsehoods. In addition, mis- and disinformation receive more attention and engagement than refutations [4], which makes it difficult to cancel out the negative consequences of false information. By testing whether fact-checks stimulate critical verification skills among recipients, we can arrive at novel insights on the potential longer-term impact of corrective information in stimulating media literacy. In other words, does exposure to fact-checks that critically scrutinize dubious arguments educate news users in how to act as fact-checkers themselves when coming across statements that raise doubt? In the remainder of this article, we will explain our methodological approach and the first preliminary findings resulting from this experiment.

Approach

We used an online survey-experiment conducted among a sample of 752 Dutch citizens. Light quotas on gender, education and age were used to ensure a sample composition that approximates the Dutch population. The sample consisted of 52% women and 38% of respondents had a college or university degree. About 36% were younger than 40 years old, 39% were between 40 and 64 years old and 25% were 65 years old or older. The key aim of the study was to arrive at a baseline assessment of the effectiveness of fact-check formats used in the BENEDMO consortium as well as the wider fact-checking community. The key outcomes were the correction of misperceptions, perceived intentions of fact-checkers, critical thinking and the intention to read fact-checks on other topics. To study this, we randomly exposed participants to fact-checks of a disputed health-related claim related to weight loss. The example was taken from actual misinformation statements related to dieting in order to ensure that our experimental set-up corresponds accurately to the real world. At the same time, this example is not directly subject to strong ideological biases or partisan responses. In the control condition of the experiment, participants did not see a fact-check. Here, we simply wanted to know the level of misperceptions people would have in the absence of a fact-check – it thus formed a baseline assessment of the level of misperceptions in our sample. In other conditions, people saw a fact-check that was formatted to resemble the style and layout of the fact-checking platform *Nieuwscheckers* (link).

Figure 1: Excerpt from a fact-check in Nieuwscheckers style as used in the experiment. All fact-checks were formatted in this style and were of a similar length, namely ca. 250-300 words.


NIEUWSCHECKERS
DE FACTCHECKERS VAN UNIVERSITEIT LEIDEN

Factcheck

De volgorde waarin je een maaltijd eet heeft geen effect op je gewicht

Een Franse biochemicus heeft naar eigen zeggen de succesformule gevonden om af te vallen zonder te diëten. Het geheim zit volgens haar in de volgorde waarin je verschillende maaltijdonderdelen eet, zoals eiwitten en koolhydraten. Helaas is het niet zo simpel: het totale aantal calorieën dat je eet is bepalend, niet de volgorde van inname.

Bewering

De volgorde waarin je een maaltijd eet heeft effect op je gewicht.

Oordeel

Onwaar.

Figure 1 shows an example fact-check where the refutation is presented in a classical style: the refutation is prominently featured in the title: “The order in which you eat a meal does not have an effect on your weight”. The false claim is repeated and refuted in the main body of the text, and an explicit verdict is provided, namely “Verdict False”. We further varied the use of explicit labels containing a verdict: Participants were either presented with the judgment that the article was ‘false/mostly false’ or saw a fact-check article that did not contain such a verdict. We also varied the level of facticity: Some participants were exposed to a fact-check that concluded that the health-related claims were mostly false, whereas others read a fact-check concluding that the claims were completely false.

Finally, some participants saw an alternative fact-check, namely the truth sandwich. The truth sandwich fact-check made the false claim less prominent in several ways: Rather than featuring

the false claim in the title, it only mentioned true information: “A healthy and balanced diet has a positive effect on your weight”. In the main body of the fact-check, it did not explicitly repeat the false claim or provide an explicit verdict. Instead, it wrapped the false claim in factually accurate information at the start and end of the fact-check article.

The variations in fact-checks resulted in a total of five experimental conditions and one control condition. An overview of all conditions can be found in *Table 1*.

Table 1: Overview of all conditions in the experiment:

		Fact-check style			
		Label	No label	Truth sandwich	No fact-check (Control)
Factuality level	False	Condition 1	Condition 2	Condition 5	Condition 0
	Mostly false	Condition 3	Condition 4		

After exposing people to a fact-check, we measured their misperceptions (the perceived accuracy of statements based on the misinformation), the perceived intention of the fact-checker, critical thinking, and the intention to engage in subsequent fact-checking of other misinformed statements. The research questions and main hypothesis are pre-registered on the Open Science Framework and can be accessed here. In the remainder of this article, we will focus on the preliminary findings with regard to the perceived accuracy of misinformation statements after being exposed to a fact-check article as well as the perceived intention of fact-checkers.

Preliminary findings

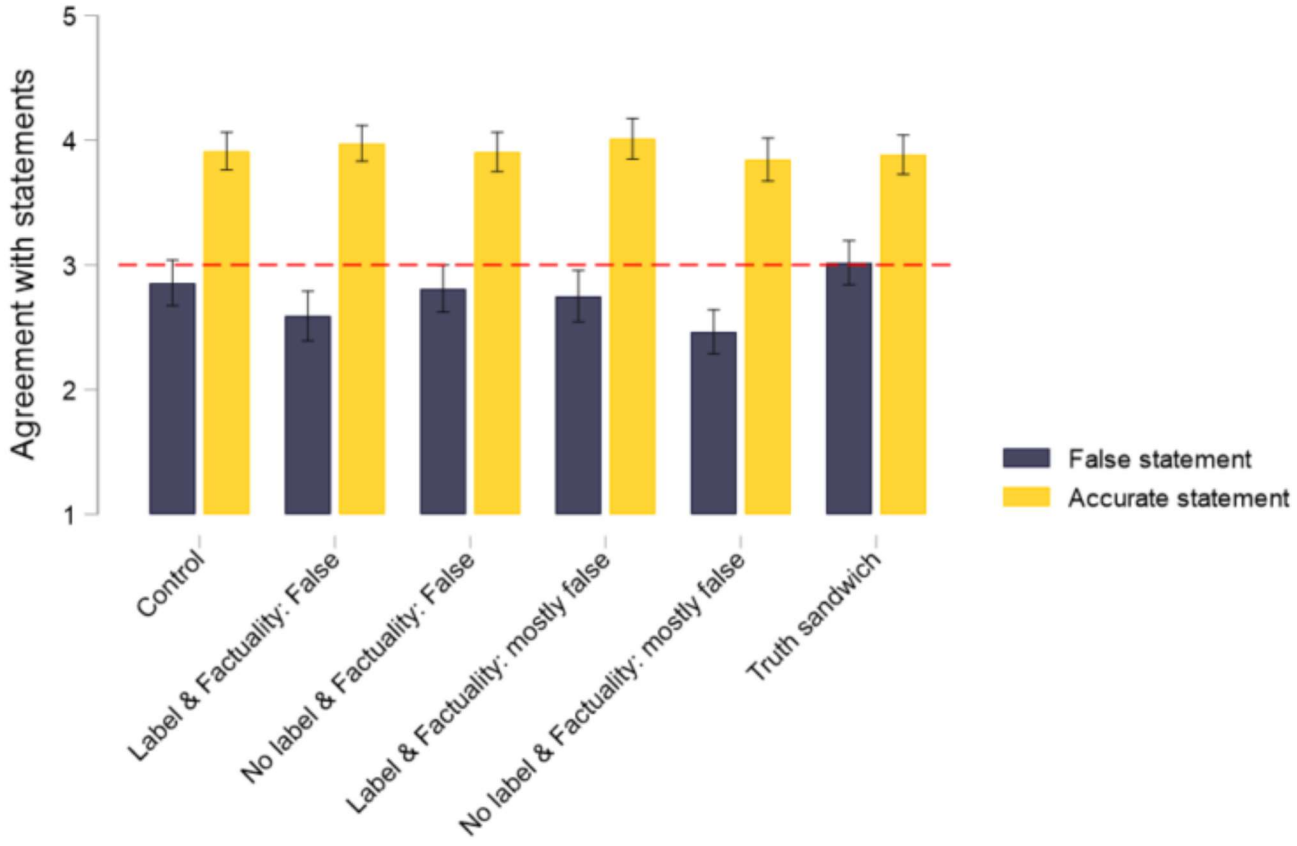
Correction of false beliefs

In our preliminary findings, we focus on how respondents evaluated one false and one accurate claim about weight loss after seeing the fact-checks. All fact-check articles that respondents read refuted the false claim that “The order in which you eat a meal has an effect on your weight”. And

each of the fact-checks concluded with the accurate claim that “The total number of calories you eat has an effect on your weight.”

Respondent evaluations of these claims were measured on five items capturing the perceived credibility, accuracy, reliability, bias and completeness of a claim. Each item of these five items was measured on a 5-point scale marking the extreme points, for example 1 = credible, 5 = uncredible. Per claim, we constructed one measure by averaging across responses to these five items. Figure 2 presents the descriptive results of respondents’ claim evaluations depending on the type of fact-check they read. It also shows a red, dashed reference line which marks the midline of the answer scale. It refers to the point where respondents evaluated the statement as neither credible nor uncredible.

Figure 2: Evaluation of accurate and false statements after reading a fact-check



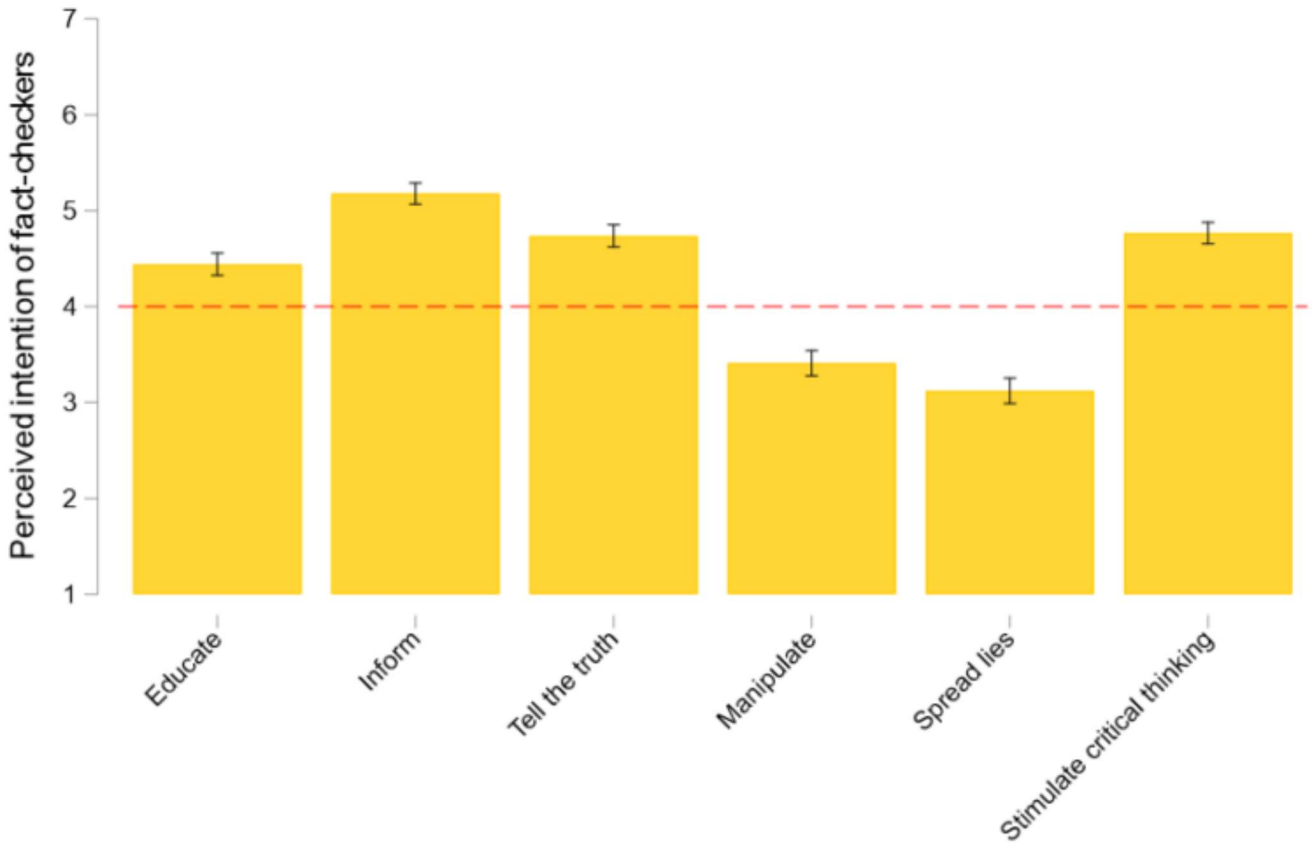
Overall, respondents judged the accurate statement to be more credible than the false statement. Zooming in on evaluations of the accurate statement, we observe that there was little variation across conditions. Compared to the control condition where respondents did not read a fact-check, respondents who did see a fact-check were not more likely to evaluate the accurate statement as credible. This is particularly surprising for the truth sandwich condition. Because the truth sandwich format highlights accurate information, we had expected that it would stimulate respondents to trust the accurate statement even more. However, this did not turn out to be the case.

In terms of evaluations of the false statement, we observe variation across conditions. While those in the control condition overall judged the false statement to be neither credible nor uncredible, our descriptive results show that respondents who did read a fact-check judged the false statement to be overall less credible. The exception is the truth sandwich condition, which again did not seem to be effective in the expected direction. The most effective fact-check was the condition that avoided using labels and that judged the statement to be “mostly false”. This condition presented the most complex fact-check: It first refuted the false statement that the order of eating has an effect on weight loss, and it then explained an indirect way in which the order of eating might have an effect, but only if it reduces the overall intake of calories. This fact-check avoided using explicit labels, which means that respondents were stimulated to draw their own conclusions. And this strategy seems to have paid off: respondents who saw this more complex fact-check were the most likely to rate the false statement as uncredible.

Perceived intentions of fact-checkers

To gain a better understanding of how participants interpreted the fact-checks, we asked them what they thought the intention of the fact-checkers was. On a 7-point scale (1 = fully disagree; 7 = fully agree) they indicated to what extent they thought that the intention of the fact-checker was to educate, inform, tell the truth, manipulate, spread lies and stimulate critical thinking. Figure 3 presents the results combined for all participants who saw a fact-check. Again, the red, dashed line presents the midpoint of the scale, which indicates that respondents neither agreed nor disagreed.

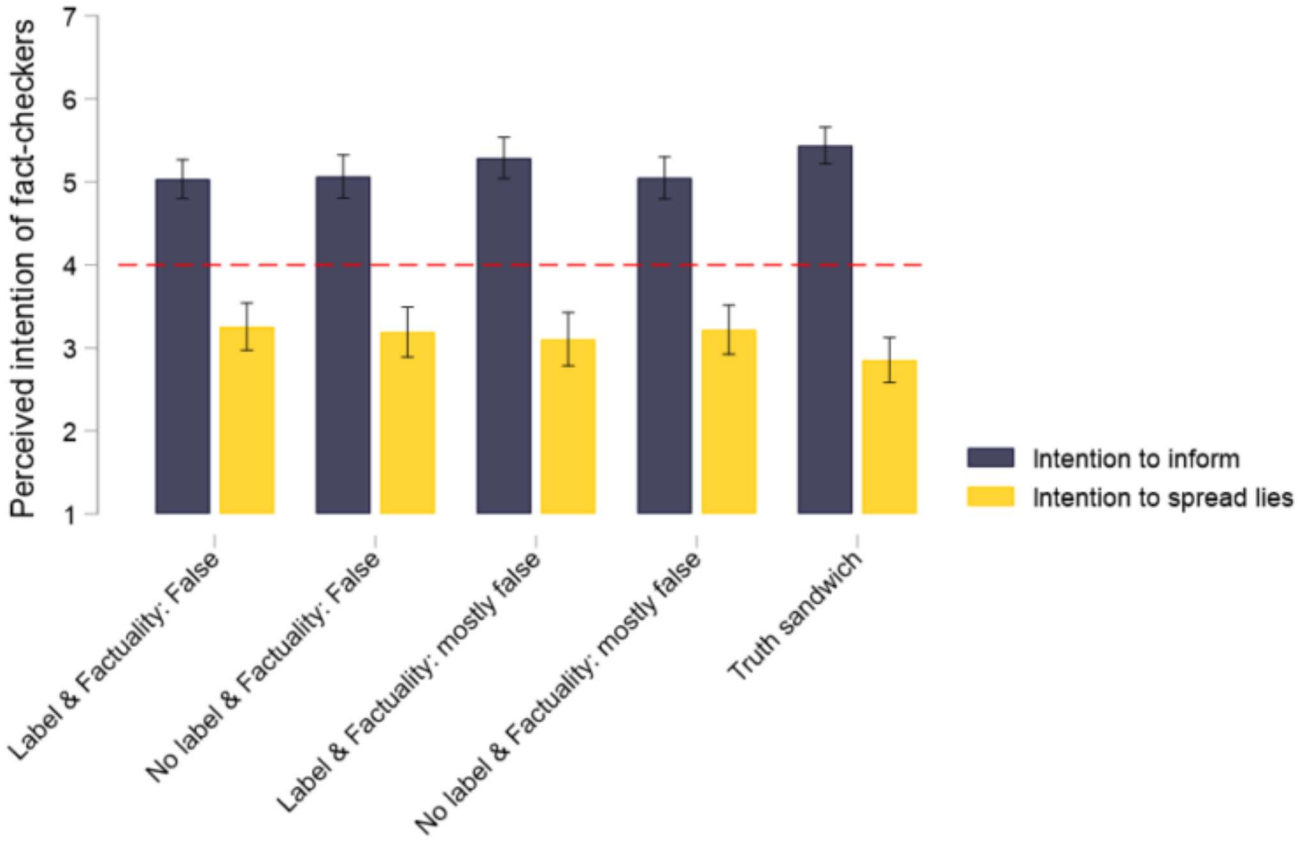
Figure 3: Perceived intentions of fact-checkers



The overall picture is positive: The intentions to inform and to stimulate critical thinking were perceived as most prominent and clearly above the midpoint. This was followed by the intention to tell the truth and to educate. Similarly, the intentions to manipulate or to spread lies were perceived as relatively unlikely.

We were also interested in how the perceived intentions varied depending on the fact-check format that respondents saw. Per fact-check format, Figure 4 presents the two intentions that were perceived to be most and least likely, namely the intention to inform and the intention to spread lies.

Figure 4: Perceived intentions of fact-checkers to inform vs. to spread lies depending on fact-check format



While the differences across fact-check formats were relatively small, this time the truth sandwich emerged as the most effective format: Our descriptive results show that respondents who read the truth sandwich format were most likely to think that the fact-checker's intention was to inform and they were also the least likely to think that the intention was to spread lies.

Conclusions

Even though the truth sandwich format did not seem to be as effective as other formats in changing beliefs related to concrete statements, it might have other, more indirect benefits, such as promoting trust in fact-checkers.

If the primary aim of fact-checking is to correct false beliefs, then our preliminary results suggest that avoiding clear labels and telling a more complex story is most beneficial in reaching this aim. When respondents saw a fact-check that avoids labels and admitted that the false claim might be true under specific circumstances, readers were more convinced. This format of presenting a fact-check might be the least likely to trigger resistance, because it presents a more complete picture and leaves room for readers to form their own opinions.

That said, the descriptive findings presented here should be judged as preliminary. Follow-up research is needed to understand whether the results of our first experiment are generalizable to other topics that are more politicized. In addition, more complex analyses are needed to better interpret the preliminary findings and to provide a more complete understanding of some of the underlying mechanisms. A next step is to analyze other outcomes of reading different fact-check formats, such as critical thinking and intentions to engage in subsequent fact-checking. Like our

preliminary results show, the impact of fact-checking depends not only on the fact-check format itself, like the use of labels or the truth sandwich format. But it also depends on what we consider the desired impact to be.

Authors: Marina Tulin, Michael Hameleers en Claes de Vreese

[1] Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350-375.

[2] Hameleers, M., & Van der Meer, T. G. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers?. Communication Research, 47(2), 227-250.

[3] König, L. M. (2022). Debunking nutrition myths: An experimental test of the “truth sandwich” text format.

[4] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146-1151.
