Debunking Misinformation in Advertising

Jessica Fong, Tong Guo, Anita Rao[*]

July 2, 2021

Abstract

The prevalence of misinformation in advertising has spurred various interested parties – regulators, the media and competing firms - to debunk false claims in the marketplace. This paper studies whether such debunking messages provided by these parties can reduce the impact of misinformation on consumer purchase behavior. If so, does debunking effectively change consumers' misbeliefs, a prediction consistent with standard Bayesian updating, or does it merely reinforce consumers' correct beliefs, a prediction consistent with confirmation bias? We design and implement a conjoint experiment that enables us to measure willingness-to-pay under exposure to real-world misinformation and debunking messages. Focusing on three ingredients in product categories where misinformation is prevalent (aluminum in deodorants, fluoride in toothpastes, and GMOs in food), we find that debunking plays an important role in mitigating the impact of misinformation. Debunking can reduce even the strongest misinformed beliefs, a promising finding for policymakers aiming to correct such misbeliefs in the marketplace. We discuss the incentives for firms to debunk or introduce new products that conform to misinformation.

[*] Jessica Fong (jyfong@umich.edu) is an Assistant Professor of Marketing at the Stephen M. Ross School of Business, University of Michigan. Tong Guo (tong.guo@duke.edu) is an Assistant Professor of Marketing at the Fuqua School of Business, Duke University. Anita Rao (Anita.Rao@chicagobooth.edu) is an Associate Professor of Marketing at the Chicago Booth School of Business. The authors are grateful to the seminar participants at Marketing Science Conference 2021, Virtual Quant Marketing Seminar, AMA Marketing and Public Policy Conference 2021, Duke University, University of Michigan, UCLA, UCSD, USC, University of Toronto, Wharton, and Washington University at St. Louis for helpful comments.

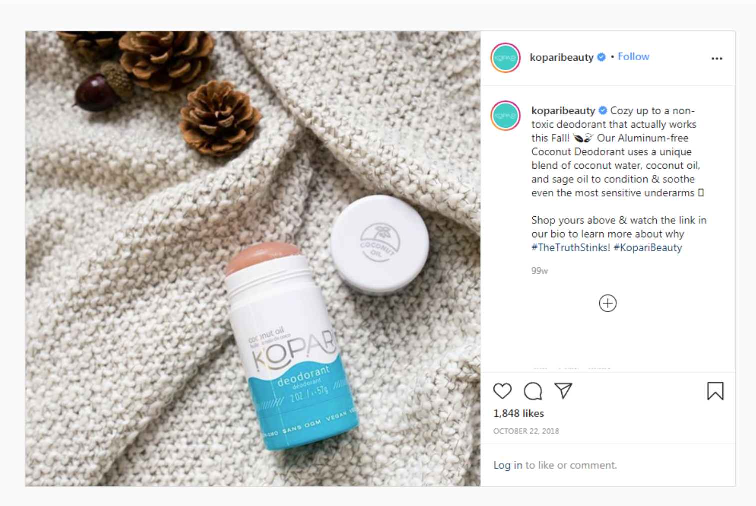
1 Introduction

Misinformation in advertisements is a widespread issue. Between 2015 and 2020 alone, the FTC filed 172 cases regarding misleading advertising and marketing, with settlements up to $191 million (FTC, 2019, 2020). Such misinformation not only harms consumers who purchase the product, but also spreads misinformation about the entire product category, creating negative spillovers into other products. For example, Kopari, a relatively new entrant in the deodorant market states their product is "aluminum-free" and is therefore non-toxic, as shown in their Instagram post displayed in Figure 1. Consumers who view this post may form new beliefs that aluminum in deodorants and antiperspirants is harmful, and may increase their willingness-to-pay for aluminum-free products, even though the claims are not supported by scientific evidence. The digital era magnifies the severity of this problem, especially in the social media context, as it is difficult to identify and remove misleading ads before it is seen by a potentially large number of consumers.

Regulators, the media, and competing firms have taken an active role in debunking such misinformation. For example, Speed Stick, on its website, highlights the lack of scientific evidence suggesting aluminum in antiperspirants and deodorants is harmful (Speedstick, 2020). However, little is known whether such debunking messages are effective and whether various sources (media vs regulator vs competitor) differ in their abilities to alter consumers' behavior led by misperceptions. Moreover, because of the fast-spreading nature of such misinformation, the extent of misinformed beliefs in the current population might be large, and it is unclear whether debunking would work on consumers with pre-existing strong (mis)beliefs.

This paper aims to understand if misinformation influences consumers' willingness-to-pay and whether debunking from various sources can effectively revert the effect of misinformation. Furthermore, this paper asks if the impact of misinformation and debunking varies by the level of consumers' existing beliefs about the focal ingredient's harmfulness.

Different theories for how individuals update their beliefs based on presented information lead to differing predictions on whether debunking would be effective for individuals with certain priors. On one hand, a standard Bayesian updating framework predicts that consumers with the most different ex-ante beliefs from the presented information change their willingness-to-pay the most. The Bayesian framework would therefore predict that those with misinformed beliefs about the ingredient increase their willingness-to-pay the most upon seeing information presented in debunking messages. On the other hand, confirmation bias literature demonstrate that individuals tend to ignore information that conflicts with their existing beliefs (Nickerson, 1998). It therefore predicts that debunking messages would



koparibeauty

koparibeauty Cozy up to a non-toxic deodorant that actually works this Fall! Our Aluminum-free Coconut Deodorant uses a unique blend of coconut water, coconut oil, and sage oil to condition & soothe even the most sensitive underarms

Shop yours above & watch the link in our bio to learn more about why #TheTruthStinks! #KopariBeauty

99w

1,848 likes

OCTOBER 22, 2018

Figure 1: Screenshot of an Instagram post by Kopari

be the least effective for consumers with strong pre-existing misinformed beliefs. Empirically measuring responses along this dimension, i.e., varying levels of existing (mis)beliefs, is important because of the different theoretical predictions which have different policy implications: one implies that debunking works for those with misinformed beliefs (an ideal outcome for a policy maker) and the other implies that debunking merely re-affirms beliefs for those without misinformed beliefs.

Empirically measuring the impact of misinformation and debunking is challenging for two reasons. First, exogenous variation in when brands present this information is rare. For example, the introduction of new products (e.g., non-GMO products) might coincide with an uptick in demand for the attribute (e.g., non-GMO), making it hard to disentangle consumer trends from message-induced demand. Second, it is almost impossible to run a field experiment in this setting because of the inherent deception involved. Debriefing those who were exposed to the deceptive ads (an IRB requirement) is practically infeasible. We therefore utilize a preference measurement tool in which we can expose consumers to various treatments and debrief them soon after. To ensure we measure the true willingness-to-pay, we use an incentive compatible choice-based conjoint to elicit preferences. Choice-based conjoint has been shown to be extremely adept at recovering and predicting consumer preferences across various attributes, leading to widespread academic and industry acceptance (Green and Rao, 1971; Green and Srinivasan, 1990). We combine the conjoint with a between-subject experimental setup, enabling us to not only recover preferences of the focal ingredient, but also measure how these preferences change under varying exposures to misinformation and debunking.

The experiment is designed to elicit preferences for various attributes including the ingredient in question (e.g., aluminum) under exposure to various treatments. Participants are first assigned to one of two advertising conditions: a control condition with an ad that highlights an attribute unrelated to the ingredient in question, or a treatment condition with an ad containing misinformation about the ingredient in question. Each participant is then randomly assigned into one of four debunking conditions: the control group sees an unrelated factoid from "How Stuff Works", and the three treatment groups see debunking messages from the media, regulator or a competing firm. By design, the debunking messages across the three sources are identical; they differ only in the source.

To understand whether consumers' ex-ante beliefs about the harmfulness of the ingredient influences their responsiveness to misinformation and debunking, we design an additional survey experiment to elicit beliefs before and after exposure to treatment. We directly elicit beliefs surrounding ingredient toxicity because inferring such beliefs from choice data alone is impossible (see Manski 2004). For example, a consumer might choose an aluminum-free deodorant driven by preferences (e.g., "I do not like aluminum because it stains clothes") or beliefs surrounding ingredient toxicity (e.g., "I do not like aluminum because it is toxic").

We define misinformation as any message that does not follow the federal law which states "an ad must be truthful, not misleading, and, when appropriate, backed by scientific evidence" (FTC). In our context, ads that state "Ingredient X is toxic" or imply that their products are "Ingredient-X-free and therefore good for you" without scientific evidence supporting the claim will be considered to be spreading misinformation. We investigate three ingredients where misinformation regarding the safety of these ingredients is widespread: fluoride, aluminum, and GMOs, with a separate survey for each ingredient.[1] For each ingredient, we select a product category that has both 1) a firm within the category that circulated messages containing misinformation, and 2) a competing firm, regulator, and media that debunks the misinformation. These criteria led to the following product categories: fluoride-free toothpaste, aluminum-free deodorant, and GMO-free nutritional shakes. To replicate the field setting to the extent possible, the ads shown in the advertising conditions are taken from actual social media posts by firms that circulated messages with misinformation, and the debunking messages are summarized from arguments presented in actual news articles, regulatory websites and competitors' websites.[2]

First, we find heterogeneity across product categories in consumers' baseline willingness-

[1] For more information about the misinformation around aluminum in deodorants, see WebMD.com (2011). Basch et al. (2019) find a high proportion of Instagram posts mentioning fluoride contain misinformation. Regarding GMOs, in 2018, 49% of Americans believe that genetically-modified foods are worse for one's health than non-genetically-modified foods (Funk et al., 2018).

[2] Ideally, the debunking messages would contain exact text from the debunking source. However, the sources' debunking messages are too long, so we summarize them for brevity in the experiment.

to-pay: consumers prefer fluoride in toothpaste but are averse to aluminum in deodorants and GMOs in food. After exposure to misinformation, the willingness-to-pay declines substantially for the preferred ingredient fluoride (a \$1.37 or 42% decrease in baseline WTP), but does not change significantly for the non-preferred ingredients: aluminum and GMOs. An absence of response to misinformation, especially for ingredients where there is a baseline negative WTP does not necessarily mean consumers are immune to misinformation. It might merely mean that consumers already have strong misinformed beliefs, and the additional exposure in the experiment does not do much to shift these beliefs. Debunking in such scenarios can therefore still play an important role.

Second, we find that debunking increases the willingness-to-pay across all three ingredients, and is able to undo the damage caused by an additional dose of misinformation. Specifically, debunking after misinformation restores the willingness-to-pay for fluoride back to the pre-misinformation level, and increases the willingness-to-pay for aluminum by \$1.21 (an increase of 63% relative to the baseline). The effect of debunking after misinformation, averaged over all sources, is positive but not statistically significant for GMOs. We find that debunking messages from regulators, media, and competing firms are effective at increasing WTP for both aluminum and fluoride, while only the competing firm is effective for debunking misinformation about GMOs. Although regulator and media messages are generally more effective than competitor messages, the differences across these sources are not statistically significant. We caveat this finding with the acknowledgment that in a real-world setting where misinformation-spreading ads are ubiquitous and debunking messages might be harder to access, the damage created by misinformation might be more severe.

Given the effects of misinformation and debunking on demand for certain product attributes, we quantify the incentives for new entrants to differentiate by spreading misinformation about ingredient X and incentives for incumbents to either debunk or introduce an Ingredient-X-free product. We show that a new entrant has a strong incentive to spread misinformation: a new fluoride-free toothpaste brand, for instance, nearly doubles its market share if it spreads misinformation about fluoride upon market entry. While incumbents can increase their market share by debunking misinformation, introducing an Ingredient-X-free product yields a greater increase in market share. This may explain why we commonly see incumbents introducing products that conform to misinformation in the marketplace (e.g., Dove and Speed Stick launching aluminum-free products in 2019, 2020, respectively).

Finally, we find that debunking is able to correct consumers' misinformed beliefs and that it is most effective for those whose prior beliefs are that the ingredient is harmful, both in terms of stated beliefs and revealed preferences. This finding is consistent with predictions from a rational Bayesian updating framework, and inconsistent with predictions from confirmation bias. This finding is encouraging for policymakers who would want their debunking messages to have an impact on the most misinformed.

The rest of the paper proceeds as follows. We review the literature and highlight this paper's contributions in Section 2. We provide a theoretical framework of how misinformation and debunking can impact beliefs in Section 3. Section 4 describes the experiment design and Section 5 describes the survey details and the data. Section 6 reports the demand estimates. We discuss our findings and conclude in Section 7.

2 Literature Review

Advertising has been theoretically and empirically well researched. The vast majority of this literature has focused on truthful advertising with deceptive advertising only recently receiving empirical attention. Recent work has focused on review fraud (Mayzlin et al., 2014; Luca and Zervas, 2016; He et al., 2020) and false claims (Rao and Wang, 2017; Avery et al., 2013; Chiou and Tucker, 2018; Rao, 2020; Kong and Rao, 2021).

Empirically measuring causal effects of false information is challenging because creating exogenous variation that spreads misinformation is not feasible: the FTC strictly prohibits such deceptive advertising, and the IRB requires debriefing anyone exposed to the ad, which might not be feasible in a field setting. Therefore, most empirical work uses a policy change that eliminates the source of misinformation, such as a regulator- or platform- induced ban (Rao and Wang, 2017; Chiou and Tucker, 2018; Rao, 2020). Although such policy changes provide exogenous variation in the amount of misinformation in the marketplace, such cases are rare. Moreover, we do not know if the effect of *removing* misinformation is symmetric to the effect of *direct exposure* to misinformation, a primary focus of our paper. We contribute to this area by directly measuring the causal effect of an additional exposure of misinformation on demand in a controlled experiment[3], which enables exogenous manipulation of debunking and debriefing.

Current efforts in combating misinformation in ads generally take one of three approaches. The first eliminates the source of misinformation via bans, shutdowns and downvotes (e.g., Chiou and Tucker, 2018; Pennycook and Rand, 2019a). However, it is difficult, if not impossible, to remove misinformation before it is seen by a potentially large audience online, especially in the social media setting. For example, a recent study found that it can take up to 22 days for the platform to downgrade and issue warning labels on Covid-related misinformation, creating 117 million exposures before cracking down (Avaaz, 2020). The second

[3] Early work in marketing has also studied the impact of exposing individuals to misinformation, but on stated purchase intentions (e.g., Olson and Dover (1978); Dyer and Kuehl (1978)).

aims to inoculate the audience against misinformation by nudging (Thaler and Sunstein, 2009). Most nudge interventions, however, involve interactions between the message sender and receiver, with the goal being to achieve reflection on the part of the consumer (e.g., Lorenz-Spreen et al. 2020; Pennycook et al. 2020) which might be harder in the context of social media. Another means of inoculation involves tagging questionable content by fact checkers. While widely adopted by platforms, recent research has shown that fact-checking tags not only can be ineffective (Guess et al., 2018) but also can backfire by causing readers to assume that non-tagged articles are true (Pennycook et al., 2020).

The third approach provides corrective messages by credible sources to directly debunk misinformation in advertising. See Walter and Murphy (2018) and Wilkie et al. (1984) for a comprehensive review of the effect of corrective information across many contexts. Most of this work, which focuses on "self-correction" by the firm (as a result of FTC lawsuits) or corrections directly from the FTC, demonstrates a small but positive effect of corrective advertising on the reduction of stated misbeliefs, both in the lab (Mazis and Adkinson, 1976; Dyer and Kuehl, 1974) and in the field (Bernhardt et al., 1986; Armstrong et al., 1983). We contribute to this literature by studying debunking messages across various sources (competitor, media and regulator) and investigating heterogeneous responses to debunking by the extent of consumers' existing (mis)beliefs in the marketplace.

Different theories on how consumers update their beliefs provide different predictions on which consumer is most impacted by the debunking message. On one hand, psychological theories suggest that corrections that are incompatible with existing (mis)beliefs tend to be processed less fluently, and are therefore less likely to be accepted (Lewandowsky et al., 2012; Nickerson, 1998). On the other hand, economic theories suggest that individuals update their beliefs in a Bayesian fashion, (Grether, 1980; El-Gamal and Grether, 1995; Holt and Smith, 2009; Coutts, 2019) suggesting corrections that are different from existing (mis)beliefs may have the highest efficacy. Based on these differing theoretical predictions, it is unclear whether a policy maker with intentions to correct the most misinformed beliefs would be successful in their goals. We therefore take this question to data by eliciting beliefs directly (Manski, 2004; Delavande, 2008; Shin et al., 2012; Schotter and Trevino, 2014) prior to treatment and measure the willingness-to-pay for consumers with varying levels of misbeliefs.[4]

[4] Our goal in this study is not to quantify the level of deviation from standard Bayesian belief updating, but to document the empirical evidence for or against the policy values of debunking, which is ex-ante ambiguous due to diverging theoretical predictions. Other studies have formally tested or modelled the typical assumptions used in belief updating theories with applications to heterogeneous price search and brand choices (Nyarko and Schotter, 2002; Charness and Levin, 2005; Jindal and Aribarg, 2021; Ching et al., 2021; Ursu et al., 2021).

Our work extends the literature on debunking misinformation in ads in the following ways. First, we measure the causal impact of corrective messaging on purchase decisions, rather than stated preferences, using an incentive compatible conjoint setting. This enables us to directly quantify the impact of debunking on demand controlling for brand and price effects. Second, we explore whether the efficacy of debunking messages varies by pre-existing beliefs, allowing us to comment on the mechanism by which individuals process the corrections to misleading claims in advertising. Third, we explore heteorgeneity of debunking by source. To our knowledge, no study has compared the effectiveness of debunking by competitors to that from regulators or mainstream media. While competitor advertising is more accessible than messages by regulators and media, it may be perceived as a competitive attack to the rival brand, therefore carrying little weight in correcting misbeliefs. Finding empirical evidence for or against the efficacy of competitor debunking has direct welfare and policy implications. Fourth, by experimentally creating variation in exposure to misinformation before debunking, we are able to evaluate whether debunking can “repair” the change in willingness to pay created through misinformation in ads.

Broadly, our work is also related to measuring consumer responses to information on nutrition and ingredient labels (e.g., Ippolito and Mathios 1990, 1995; Dhar and Baylis 2011; Bollinger et al. 2011; Liaukonyte et al. 2013; Hobin et al. 2017; Fernbach et al. 2019; Scott and Rozin 2020). We contribute to this literature by studying the impact of false information pertaining to various ingredients. Moreover, our paper focuses on the immediate outcomes after exposure to misinformation and debunking, especially meaningful in the online setting and social media context because platforms can precisely target consumers at the point of purchase and deliver messages when it is the most consequential.[5] Finally, our work is related to the voluminous literature on misinformation in the health, journalism and political domain.[6]

3 Theoretical Framework

In this section, we describe a framework of how ads and debunking influence beliefs, and why debunking effectiveness can vary both across debunking sources and individuals. This model also illustrates the differing predictions in beliefs when consumers update in a rational

[5] For work on the long-term effects of misinformation and debunking see Skurnik et al. (2005); Schwarz et al. (2016); Schwarz and Jalbert (2020); Hovland and Weiss (1951); Pratkanis et al. (1988).

[6] See Flynn et al. 2017; Lazer et al. 2018; Vosoughi et al. 2018; Guess et al. 2019; Pennycook and Rand 2019b; Bago et al. 2020; Dias et al. 2020; Guess et al. 2020; Linvill and Warren 2020; Pennycook et al. 2020; Walter et al. 2020; Martel et al. 2020; Simonov et al. 2020; Bursztyn et al. 2020; Barrera et al. 2020; Chen et al. 2020; Pennycook et al. 2021 for recent applications.

Bayesian fashion and when they update with confirmation bias. We model a consumer's decision to purchase a product j, where the decision depends on whether the product contains the ingredient that is deceptively advertised. To formalize this model and illustrate how beliefs about the ingredient directly affect consumers' utility, let consumer i's utility for purchasing product j be

$$u_{ij} = \beta_i x_j + Z_j \gamma + \epsilon_{ij} \tag{1}$$

where $x_j \in \{0, 1\}$ indicates whether j contains the focal ingredient, Z_j is the vector consisting of the product's other attributes (such as brand, price, packaging, flavor, scent, etc), and ϵ_{ij} is an idiosyncratic shock that varies at the consumer-product level, assumed to be i.i.d. The coefficient of interest is β_i, i.e., i's preference for the focal ingredient. For example, if the product is deodorant, then x indicates whether the deodorant contains aluminum, and β_i is i's preference for aluminum in deodorant.

Advertisements containing (mis)information and debunking messages about the focal ingredient can change i's preference by changing i's beliefs about whether the ingredient is harmful to one's health. Specifically, let the data generating process for β_i be

$$\beta_i = \tau + \delta\theta_i + \eta_i \tag{2}$$

where $\theta_i \in [0, 1]$ is i's belief that the ingredient is harmful. δ represents the population average preference for harm, which is most likely negative. τ is the population average inherent preference for the ingredient, which can be either positive or negative.[7] η_i are individual-level differences in the preferences for the ingredient that are unexplained by θ_i.

Below we describe a sequence of events that leads to θ_i. Let H denote the true state of the world where the ingredient can be "harmful" or "not harmful", and let the information in the message content be denoted by *claim*, where *claim* = `not harmful` represents the scenario when the message claims "the ingredient is not harmful", and *claim* = `harmful` when the message claims "the ingredient is harmful". The timing is as follows:

1. Nature decides $H = \{\text{Harmful, Not Harmful}\}$.

2. i has a prior belief that the ingredient is harmful, denoted by θ_i^0.

3. i can be exposed to an ad that claims that a specific ingredient is harmful (*claim* = `harmful`). If she sees this ad, i then updates her belief about ingredient toxicity to θ_i^{ad} after viewing the ad. If she does not see this ad, her belief remains θ_i^0.

[7] For example, consumers may prefer deodorant with aluminum because aluminum in deodorant is effective at preventing sweat buildup. Consumers may prefer fluoride in toothpaste as it is the key ingredient to prevent cavity. They may prefer nutrition shakes made from GMO crops because GMO crops generates smaller environmental impact.

4. i can then be exposed to a debunking message. If she is exposed to debunking, she updates her belief to $\theta_i = \theta_i^{ad,d}$ if she previously saw the ad, and $\theta_i = \theta_i^{0,d}$ if she did not see the ad prior to debunking. Similarly, if she is not exposed to debunking, then $\theta_i = \theta_i^{ad}$ (if she sees the ad) or $\theta_i = \theta_i^0$ (if she did not see the ad).

Note that θ_i^0 is a Bernoulli prior, where i's uncertainty about H is represented by how far θ_i^0 is from 0 or 1 (e.g., if $\theta_i^0 = 0.5$, i is uncertain whether x is harmful). We abstract away from the distribution of an individual's prior as in Manski (2004), where θ_i^0 is sufficient to provide a probabilistic summary of ambiguity. This setup is similar to the framework used in the Bayesian Persuasion literature (Kamenica and Gentzkow, 2011).

The extent that i updates her posterior in response to an ad with misinformation or a debunking message depends on i's evaluation of the "trustworthiness" of the source making the harmfulness claims. We define trustworthiness in the following way. Let the trustworthiness of a source be $\pi \in [0, 1]$. If the source is completely trustworthy (i.e., $\pi = 1$), then the source always truthfully reports. Formally, trustworthiness can be defined by the following:

$$p(claim = \texttt{harmful}|H = \texttt{harmful}) = \pi \tag{3}$$

$$p(claim = \texttt{not harmful}|H = \texttt{harmful}) = 1 - \pi \tag{4}$$

$$p(claim = \texttt{not harmful}|H = \texttt{not harmful}) = \pi \tag{5}$$

$$p(claim = \texttt{harmful}|H = \texttt{not harmful}) = 1 - \pi \tag{6}$$

Equation 3 states that if the ingredient is harmful, then the source claims that the ingredient is harmful with probability π. Similarly, Equation 5 states that if the ingredient is not harmful, the source claims it is not harmful with probability π. Equation 5, therefore, captures our notion of debunking where in all instances we study the focal ingredient is not harmful. Conversely, Equation 6 states if the ingredient is not harmful, then the source claims that it *is* harmful with probability $1-\pi$. Equation 6, therefore, captures our definition of misinformation. We denote consumers' evaluation of the trustworthiness of the toxicity message in the ad as π^a. The evaluated trustworthiness for a debunking message from a source is denoted by π^d.

3.1 Bayesian Updating

According to Bayes' rule, the posterior belief that the ingredient is harmful after receiving information $claim = \texttt{harmful}$ from a source with trustworthiness π, given a prior θ_i^0, is the

Table 1: Posterior Beliefs under Bayes' Rule

Sees Ad with Misinformation?	Sees Debunking Message?	Posterior
No	No	θ_i^0
Yes	No	$\frac{\pi^a\theta_i^0}{\pi^a\theta_i^0+(1-\pi^a)(1-\theta_i^0)} = \theta_i^{ad}$
No	Yes	$\frac{(1-\pi^d)\theta_i^0}{(1-\pi^d)\theta_i^0+\pi^d(1-\theta_i^0)} = \theta_i^{0,d}$
Yes	Yes	$\frac{(1-\pi^d)\theta_i^{ad}}{(1-\pi^d)\theta_i^{ad}+\pi^d(1-\theta_i^{ad})} = \theta_i^{ad,d}$

Notes: The posterior beliefs about whether the ingredient is harmful, given the ad and debunking message the consumer receives. π^d is the trustworthiness of the debunking message and π^a is the trustworthiness of the harmfulness message in the ad regarding ingredient. θ_i^0 is the prior belief that x is harmful, before the consumer has viewed any ad or debunking message.

following.

$$\begin{aligned}
\theta_i^{post} &= p(H = \texttt{harmful}|claim = \texttt{harmful}) && (7)\\
&= \frac{p(claim = \texttt{harmful}|H = \texttt{harmful})p(H = \texttt{harmful})}{p(claim = \texttt{harmful})} && (8)\\
&= \frac{\pi\theta_i^0}{\pi\theta_i^0 + (1-\pi)(1-\theta_i^0)} && (9)
\end{aligned}$$

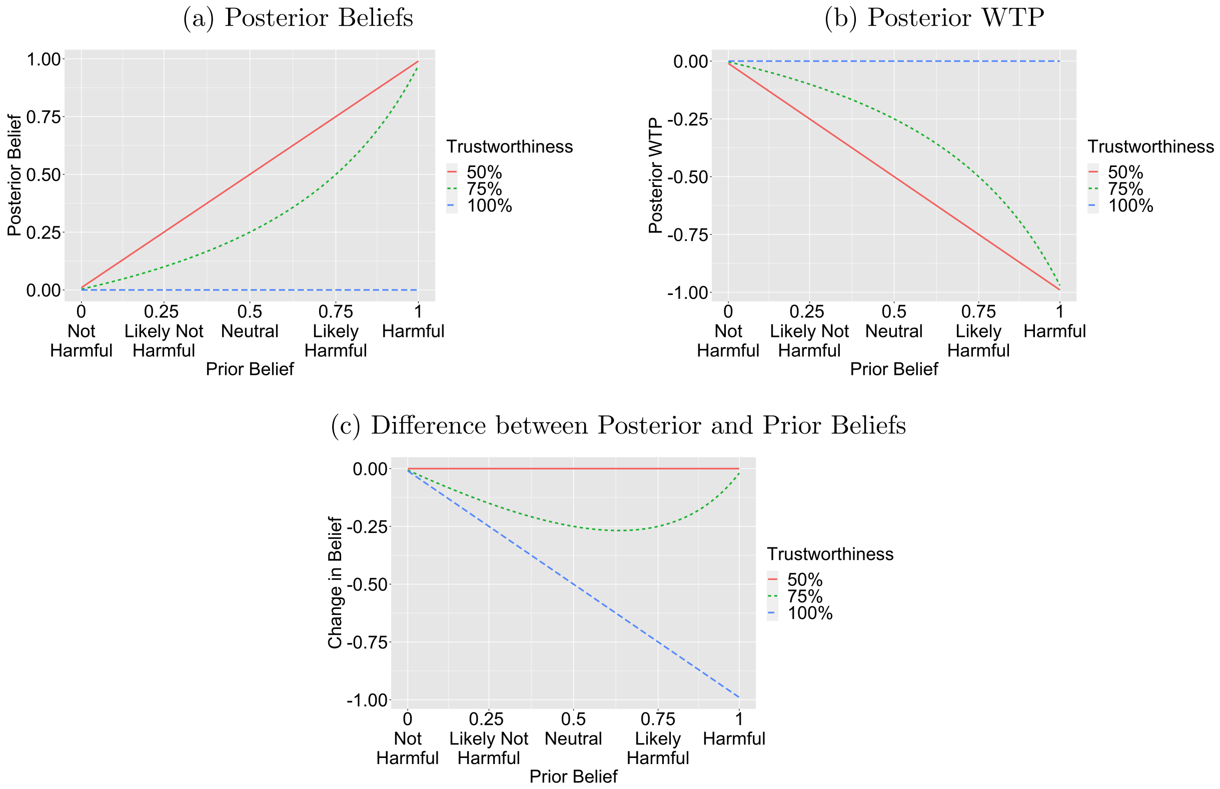
Given that individuals may be exposed to misinformation and/or debunking, this creates four distinct sequences of events that lead to different posterior beliefs and willingness-to-pay for the focal ingredient. The posterior beliefs about whether the ingredient is harmful given misinformation and/or debunking are displayed in Table 1.

To illustrate how information impacts beliefs, Figures 2a and 2b illustrate the posterior belief and willingness-to-pay after an individual is exposed to a debunking message ($claim = \texttt{not harmful}$), based on her prior belief and the debunking source's trustworthiness. Her WTP is a linear transformation of her preference for the ingredient (β) and is given by:

$$WTP = \frac{\tau + \delta\theta^{post}}{|\gamma|} \qquad (10)$$

These figures depict scenarios when $\pi \in [0.5, 1]$ which are of interest: $\pi = 0.5$ when the signal from the source is noisy and is of no informational value, $\pi = 1$ when the information is completely trustworthy. Values below 0.5 occur when consumers interpret the source's message in a directly opposite manner and are not relevant to our setting (e.g., consumers see a message stating aluminum is harmful and believe that aluminum is not harmful).

Figure 2: Bayesian Posterior Beliefs and WTP After Debunking



(a) Posterior Beliefs

(b) Posterior WTP

(c) Difference between Posterior and Prior Beliefs

Notes: Figures 2a and 2b displays θ^{post}, an individual's posterior belief that the ingredient is harmful, and her corresponding WTP, respectively, for the ingredient after she sees a debunking message for varying levels of the debunking source's trustworthiness. The posterior belief is derived from the equations in Table 1, and the WTP is derived from the posterior belief at the following parameter values $\gamma = 1$, $\tau = 0$, and $\delta = -1$.

The figures illustrate the intuition that the more trustworthy the debunking source is, the closer the posterior belief is to 0 (i.e., the ingredient is Not Harmful). Similarly, the more trustworthy the debunking source is, the higher the WTP the individual has for the ingredient. Figure 2c depicts the change in beliefs relative to the prior after receiving a debunking message. It indicates that if individuals update their beliefs according to Bayes' rule after seeing a trustworthy message, generally, the update is larger for those who believe the ingredient is more likely harmful. However, when the source is not 100% trustworthy, the level of updates becomes non-monotonic with regard to the prior belief: debunking will be less effective for those with strong priors that the ingredient is harmful ($\theta^0 = 1$) compared to those who are more uncertain ($\theta^0 \in [0.5, 1)$). Nonetheless, even in this scenario with lower levels of trustworthiness, those who consider the ingredient to be harmful ($\theta^0 \in [0.5, 1)$) are the ones who update their beliefs the most, as evidenced by the right-skewed nature of the curve. From a policy maker's perspective, such a change in beliefs is ideal because the debunking message corrects beliefs for those with the most misinformed beliefs, i.e., for those who think the ingredient is harmful.

3.2 Confirmation Bias

In direct contrast to rational Bayesian updating, confirmation bias would predict that those with strong misbeliefs would ignore debunking messages. Similarly, those who think the ingredient is not harmful would not be swayed by misinformation. These patterns arise because confirmation bias suggests that users select information that agrees with their beliefs and ignore any contradicting information. We illustrate these predictions using a specific model of confirmation bias in Appendix A. In particular, the curve is left-skewed (shown in Figure A2b, as opposed to the right-skewed curve depicted in Figure 2c) and captures the intuition that according to confirmation bias, those whose beliefs are consistent with the content of the debunking message are willing to update, whereas those whose beliefs are contradictory to the content of the debunking message are less willing to update. From a policy maker's perspective, this scenario is less promising because those with the most misinformed beliefs are less likely to alter their beliefs upon viewing messages with contradictory, albeit true, information.

4 Experiment Design

Measuring the effectiveness of debunking methods is challenging in the field setting, as it is difficult to separate the timing of ads containing misinformation and debunking messages from general changes in public opinion and demand using observational data. Additionally, field experiments are also not feasible due to the deceptive nature of the ads. Therefore, we implement an incentive-compatible choice-based conjoint experiment that separately measures the effect of misinformation in advertising and the effect of debunking on consumer preferences and demand. More specifically, we measure whether debunking messages impact demand, how this effect varies by the source of the debunking message, and whether debunking can "undo" the demand effects resulting from misinformation.

Three decisions had to be taken involving 1) the choice of product categories, 2) implementation of treatment conditions and 3) the method of capturing the outcome of interest, i.e., purchase. We go over the details informing each of these decisions in the following sections 4.1-4.3. Section 4.4 describes how we design the belief elicitation survey.

4.1 Choice of Product Categories

We identified categories where certain products market themselves as containing "ingredient-X-free", and either directly state or indirectly imply that ingredient X is toxic. Moreover, ingredient X is a prominent ingredient in almost all products in that category. We further restricted attention to categories where there are debunking messages from competitors, media and regulators.

These criteria helped us identify three product categories and the ingredient in question: 1) deodorants and aluminum, 2) toothpastes and fluoride, 3) nutritional shakes and GMO. These ingredients remain controversial in the US, despite no scientific evidence of harm.[8] In the deodorant category, Kopari states that their deodorants are aluminum-free and implies that other deodorants are toxic. See Figure 4b for an example. Competitors (Speed Stick), the media (MSN) and regulators (CDC) have provided information to consumers that aluminum, when used topically, is safe for healthy individuals. In the toothpaste category, Risewell highlights their toothpastes are fluoride-free and encourages consumers to avoid the toxic ingredients found in traditional fluoride toothpastes (Figure 5b), while competitors (Colgate), the media (NBC News) and the CDC have all highlighted why fluoride is beneficial and how fluoride-free toothpastes can actually harm oral health.[9] In nutritional shakes, Orgain highlights that their products are GMO-free and therefore contain only the "good stuff", implicitly stating that products with GMOs are bad (Figure 6b). Competitors (Soylent), the media (NBC) and the regulator (FDA) have pointed out that genetically modified plants are not only safe to consume but can actually be beneficial to the environment. Table 2 reports the category, product/firm making the false claims and the debunking sources used in this study.

Table 2: Categories, Products Making False Claims, and Debunking Sources

	Debunking Source		
Category (Product)	Firm	Media	Regulator
Deodorant (Kopari)	Speed Stick	MSN News	CDC
Toothpaste (Risewell)	Colgate	NBC News	CDC
Shakes (Orgain)	Soylent	NBC News	FDA

[8] See Penn Medicine (2019), CDC (2020), and Feldmann et al. (2000) for more information about myths and truths about aluminum, fluoride, and GMOs, respectively.

[9] Avoidance of fluoride is a unique US phenomenon: there is more controversy around fluoride in the US relative to Europe, where consumers "want fluoride in their toothpastes, because they know it's one of the most important ingredients for rebuilding tooth decay" (Glossy.com, 2020). For more information about fluoride, see AFS (2021).

4.2 Treatment

Figure 3 displays the between-subject experimental design, in which survey participants are randomized into an ad group and a debunking group. All ads and debunking messages are displayed as Tweets, because the products predominantly promote their products through social media.

Participants are first randomized into receiving either a control ad or a treatment ad. Participants randomized into the treatment ad group receive an ad that contains misinformation about the focal ingredient, while those randomized to the control group receive an ad for the same product that does not mention the focal ingredient. The ads in both groups are actual content from the company. Although not tagged as ads on the platform, we refer to such content as ads because they are messaging that highlights the firm's products, is aimed at consumers, and comes directly from the brand.[10] Figures 4 - 6 display the control and treatment ads across all three products.[11]

Participants are then randomized into one of the following debunking sources: control, competitor firm, media, or regulator. For a given debunking source, with the exception of the control group, the participant sees a Tweet from the source that debunks the notion that the focal ingredient in the product is toxic. For example, in the toothpaste/fluoride survey, a participant in the competitor firm, media, or regulator group sees the message "Fluoride-containing toothpastes are safe. Fluoride in toothpastes prevents cavities and scientific studies have found no conclusive evidence that it causes adverse health effects.". The message is accompanied by a link that leads to a real article from the source with the same overall message. These debunking messages were taken from actual articles across all sources. To ensure that only the source varies across all treatment arms, we hold the content of the debunking message constant. Figure 7 presents an example of the debunking message, as seen by the participant. After the ad and treatment exposures, we conduct verification checks to check whether the participant can recall the source of the ad and the debunking message.

The control debunking message is a factoid about the product category that contains

[10] The control ad, which is explicitly chosen such that it does not include misinformation, is an actual firm-broadcast message and highlights another attribute (such as scent or flavor) that is orthogonal to the preference for the focal ingredient (i.e., does not interfere with the measurement of the WTP for the focal ingredient). A control ad which just excludes the misinformation would not only be a much shorter message but would also be asymmetric to the treatment ad in that it would not highlight any attribute.

[11] Not all firms' promotional messages were on Twitter: some were released on Facebook or Instagram. We are agnostic about the social media platform. To ensure that our experiment does not vary the platform of the advertisement, we decided to use Twitter as the consistent platform across all products because 1) Facebook was facing controversies in 2020 and 2) after Facebook, Twitter is reported to be the most popular social media platform for text-heavy news consumption. We also ensure other aspects of the ad, such as the picture, timestamp, and likes and retweets for both ads, are identical across control and treatment conditions.

no information about the focal ingredient. This factoid was presented as a Tweet from the website "How Stuff Works". For the remainder of this paper, we refer to the control debunking group as the No Debunk group. Table B1 in the Appendix displays the debunking messages for all products and treatment groups. Table 2 reports the sources for each product and debunking type. Note that ideally, the source would be the same across all products. However, to ensure external validity to the extent possible, the criteria for selecting the source was that the source must have an actual article that debunks the misinformation about the focal ingredient. For instance, we were unable to find an article from the CDC that debunks misinformation about GMOs, and therefore, used an article from the FDA instead.

Figure 3 summarizes the experimental setup, which is a 2x4 design. This design allows us to measure the misinformation and debunking effects separately. Misinformation effects are quantified by comparing measured preferences for the focal ingredient between the "Control Ad + No Debunk" and "Misinformation Ad + No Debunk" groups because participants in these two groups are not exposed to debunking and differ only by the ad content they were exposed to. Debunking effects for a given source are measured by comparing those in the "Misinformation Ad + Debunk" group for the given source with the "Misinformation Ad + No Debunk". This design also allows us to measure whether debunking is effective for participants who were not exposed to misinformation in this survey, but perhaps already had existing misconceptions about the ingredient prior to the survey. If debunking works even without exposure to misinformation, this will be evidenced in the difference between the "Control Ad + Debunk" and "Control Ad + No Debunk" groups.

Towards the end of the survey, we conduct further verification checks by asking participants to recall content of both the ad and the debunking message. These verification checks are placed after the conjoint questions, instead of immediately after treatment, to avoid mistakenly "treating" the respondents by the choices presented in these verification checks. Lastly, we debrief the participants after survey completion.

Figure 3: Experiment Design: Flow of Treatment and Various Treatment Arms

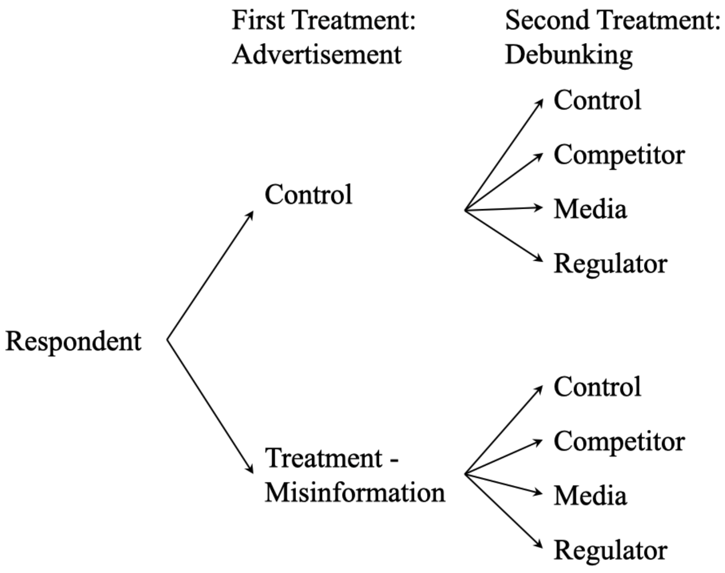
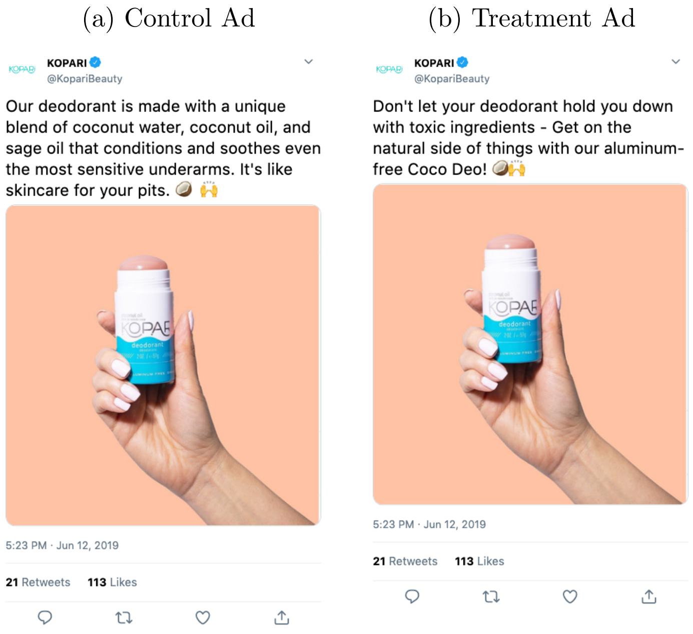


Figure 4: Ads in Aluminum Survey


(a) Control Ad

KOPARI
@KopariBeauty

Our deodorant is made with a unique blend of coconut water, coconut oil, and sage oil that conditions and soothes even the most sensitive underarms. It's like skincare for your pits. 🥥 🙌

5:23 PM · Jun 12, 2019

21 Retweets 113 Likes

(b) Treatment Ad

KOPARI
@KopariBeauty

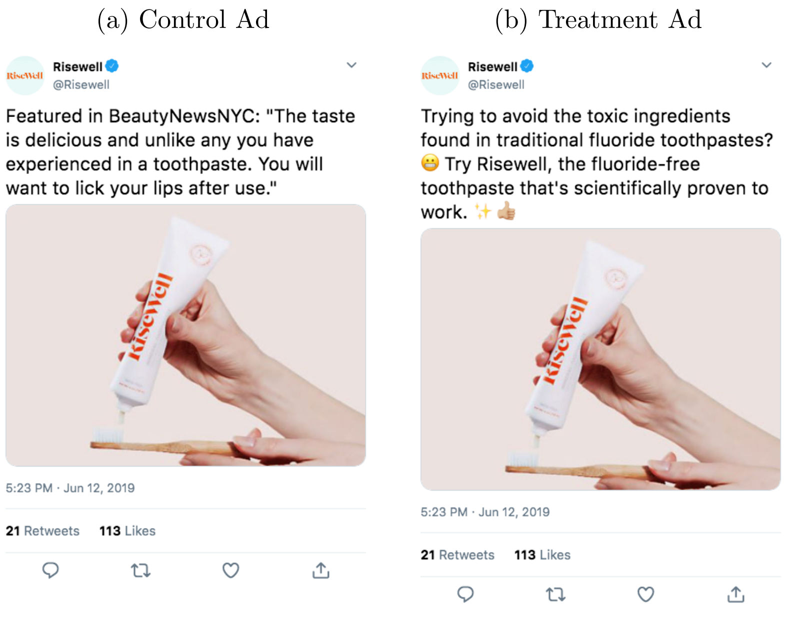
Don't let your deodorant hold you down with toxic ingredients - Get on the natural side of things with our aluminum-free Coco Deo! 🥥🙌

5:23 PM · Jun 12, 2019

21 Retweets 113 Likes

Notes: An additional treated ad shown in Appendix, Figure B1 was also used. Because the results across the two treatment ads did not differ we collapse them into one for analyses/description.

Figure 5: Ads in Fluoride Survey



(a) Control Ad

Risewell
@Risewell

Featured in BeautyNewsNYC: "The taste is delicious and unlike any you have experienced in a toothpaste. You will want to lick your lips after use."

5:23 PM · Jun 12, 2019

21 Retweets **113** Likes

(b) Treatment Ad

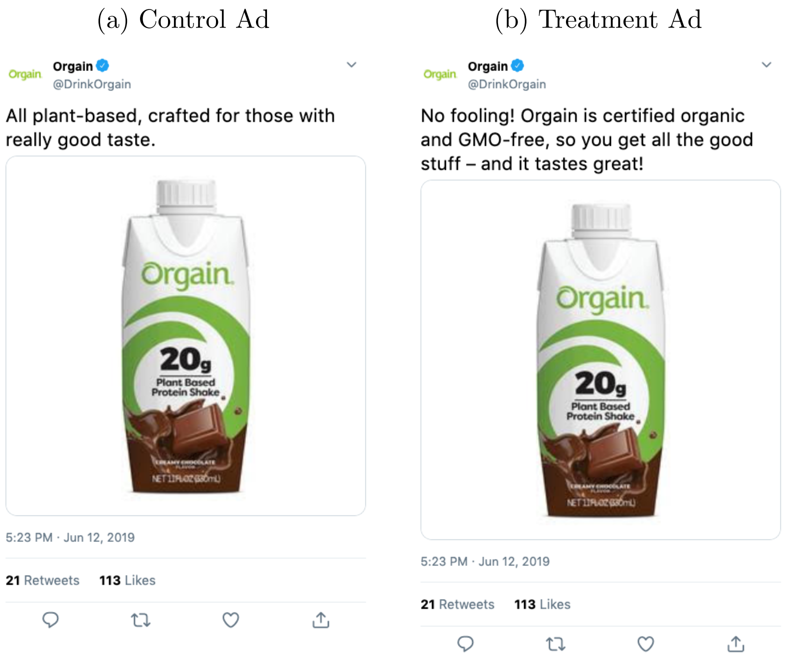
Risewell
@Risewell

Trying to avoid the toxic ingredients found in traditional fluoride toothpastes? 😬 Try Risewell, the fluoride-free toothpaste that's scientifically proven to work. ✨👍🏼

5:23 PM · Jun 12, 2019

21 Retweets **113** Likes

Figure 6: Ads in GMO Survey



(a) Control Ad

Orgain
@DrinkOrgain

All plant-based, crafted for those with really good taste.

5:23 PM · Jun 12, 2019

21 Retweets **113** Likes

(b) Treatment Ad

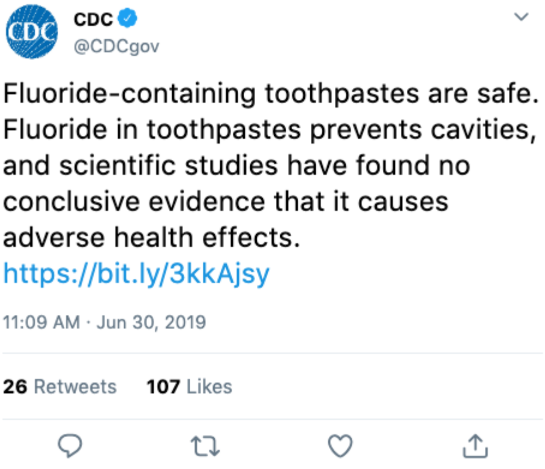
Orgain
@DrinkOrgain

No fooling! Orgain is certified organic and GMO-free, so you get all the good stuff – and it tastes great!

5:23 PM · Jun 12, 2019

21 Retweets **113** Likes

Figure 7: Example Debunking Message



CDC
@CDCgov

Fluoride-containing toothpastes are safe. Fluoride in toothpastes prevents cavities, and scientific studies have found no conclusive evidence that it causes adverse health effects. https://bit.ly/3kkAjsy

11:09 AM · Jun 30, 2019

26 Retweets 107 Likes

Notes: This is the debunking message from the regulator for the fluoride survey.

4.3 Incentive Compatible Conjoint Design

To measure consumer preferences, we designed an incentive compatible conjoint survey. See Green and Srinivasan (1978, 1990) for an overview of the conjoint literature and Ding et al. (2005) for a discussion on incentive-aligned conjoint analysis.

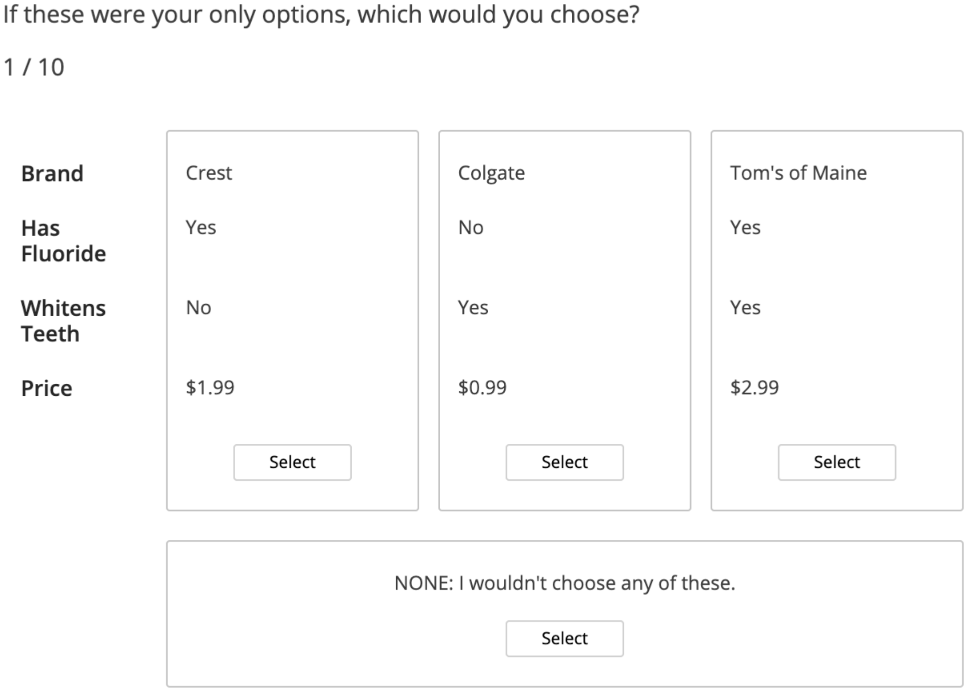
After exposure to the two treatment conditions, participants are presented with 10 conjoint choice tasks. Participants are asked to choose a product from 3 options, or none of the options. The products are unique combinations of four attributes: brand, whether it contains the focal ingredient, price, and a balancing attribute ("Whitening" for toothpaste; "Scented" for deodorant; and "Flavor" for nutritional shakes). Figure 8 provides an example of the choice task faced by participants in the fluoride survey. Table 3 details the product attributes used in the conjoint.

To ensure that the conjoint elicits participants' true preferences, the conjoint is designed to be incentive compatible. Participants are told that they have a 1-in-20 chance to win a bonus worth \$10,[12] which includes one of the products they selected. If they are selected to win the product, they receive the product that they selected for the given price and the remaining \$10-(selected price) as additional payment. For example, if a participant wins the bonus and had selected a Crest toothpaste with whitening and fluoride for \$0.99, he receives a Crest toothpaste with whitening and fluoride for \$0.99 and the remaining \$9.01 as an additional cash payment.

After the conjoint choices, we collected information about participants' usage of the

[12] For the GMO survey, this was 1-in-20 chance to win a bonus worth \$20 to accommodate the higher price of the dozen-pack of nutritional shakes.

Figure 8: Example choice task (Fluoride survey)



If these were your only options, which would you choose?

1 / 10

Brand	Crest	Colgate	Tom's of Maine
Has Fluoride	Yes	No	Yes
Whitens Teeth	No	Yes	Yes
Price	$1.99	$0.99	$2.99
	Select	Select	Select

NONE: I wouldn't choose any of these.

Table 3: Product Attributes

Product	Brand	Has Ingredient	Price	Other Attribute
Deodorant	Dove, Speed Stick, Kopari*	Has Aluminum: Yes/No	1.99, 2.99, 3.99	Scented: Yes/No
Fluoride	Colgate, Crest, Risewell*, Tom's of Maine	Has Fluoride: Yes/No	0.99, 1.99, 2.99	Whitens Teeth: Yes/No
Nutritional Shakes	Ensure, Orgain*, Soylent	GMO-free: Yes/No	1, 1.25, $1.50^{\dagger}$	Flavor: Chocolate/Vanilla

Notes: This table displays conjoint attributes for all products.
*: Denotes the advertised brand
† : Price per bottle. Due to logistics of reward distributions, lottery winners in the shake/GMO survey received a dozen shakes, so the conjoint selections are for a dozen pack of shakes.

product, their opinions about the focal ingredient, and demographics. Tables B2-B5 in the Appendix list all the questions asked in this section of the survey. Every participant is debriefed with verified scientific content after survey completion. For the remainder of this paper, we refer to these 3 surveys (aluminum, fluoride, and GMO) as the "ingredient surveys".

4.4 Eliciting Beliefs

As discussed in the literature review and illustrated in Section 3, confirmation bias and standard Bayesian updating produce different predictions on how different consumers react to misinformation and debunking depending on their existing beliefs about the ingredient. Measuring responses as a function of beliefs is important from a policy maker's perspective: Does debunking have the most impact on consumers with the most misinformed beliefs, a prediction consistent with standard Bayesian updating and an ideal outcome for a policy maker? Or does debunking have little to no impact on those with the most misinformed beliefs, a prediction consistent with confirmation bias? Using preferences in the control group as a metric for beliefs, while feasible, has the typical preferences vs. beliefs confound: a consumer might dislike aluminum (preference) but believe it is nontoxic (belief), making preference a poor proxy of belief. Manski (2004) suggests directly eliciting beliefs from respondents to circumvent this issue.

We therefore design another survey which elicits consumers' beliefs about ingredient toxicity before they respond to the choice questions.[13] For the remainder of this paper, we refer to this survey as the "beliefs survey". To avoid creating a demand effect (in which consumers see belief questions about the ingredient and infer this survey has something to do with ingredient toxicity and change their responses to satisfy the researcher), we ask beliefs questions for all other attributes included in the survey. Following probabilistic elicitation of beliefs suggested by Manski (2004), participants are asked to respond to "What do you think is the percent chance that <attribute> is harmful to your health?" with the following possible options: 0-20% (Definitely not harmful); 20-40% (Likely not harmful); 40-60% (Not sure, either way); 60-80% (Likely harmful); 80-100% (Definitely harmful).[14] We collect both the participants' prior beliefs and the reasoning behind their beliefs before treatment. At the end of the survey, we again elicit participants' posterior bliefs about ingredient toxicity. We report survey details in Section 6.1.

[13] We designed this additional survey, instead of eliciting beliefs in the ingredient surveys, because the sample size needed to detect statistically significant differences across sources and beliefs is not monetarily feasible.

[14] As mentioned in Section 3, we treat belief that "<ingredient> is toxic" as a Bernoulli prior and that uncertainty is represented by how far the consumer's response is is from 0 or 1. We abstract away from uncertainty around each probabilistic measure.

5 Survey Implementation and Data Description

5.1 Implementation

The surveys were distributed through Prolific, an online platform for survey administration and data collection. The ingredient surveys were launched sequentially in September and October 2020.[15] Such a sequential launch enabled us to exclude participants who had already taken any previous survey. We explicitly did so to avoid any possibility of familiarity with the survey and treatment conditions. The beliefs survey was launched in March 2021.

The participant pool for each survey are limited to those in the United States and those who did not participate in any of the other surveys in this study. Participants received $1.50 USD for survey completion. As previously described in Section 4.3, to ensure incentive compatability in the conjoint, participants also had a chance to win an additional bonus, which includes a product they selected in the conjoint and additional payment.

5.2 Data

In this section, we describe the data for the ingredient surveys. Discussion on the beliefs survey data is in Section 6.1.

Table 4: Sample Size for Each Treatment Group

Ad	Debunk	Aluminum*	GMO	Fluoride
Control	Control	155	171	401
Control	Competitor	146	217	407
Control	Media	140	195	405
Control	Regulator	161	204	406
Treatment	Control	296	215	411
Treatment	Competitor	292	190	386
Treatment	Media	300	187	379
Treatment	Regulator	307	180	407

Notes: *: For the aluminum survey, we used two versions of treatment ads – a strong and mild version of misinformation. We group them together into one treatment group in our main analysis, resulting in a total of approximately 300 participants in the treated ad-debunking groups. That is, 150 participants in the control debunking group saw a control ad, 150 saw the strong version of the treatment ads, and 150 saw the mild version of the treatment ad. We report the milder version of the treatment ad and results in Appendix C.

In total, 6,558 individuals completed the ingredient surveys, with 1,797 participants in the aluminum survey, 3,202 in the fluoride survey, and 1,559 in the GMO survey. The

[15] The surveys were launched on three Wednesdays: September 9, September 23, October 21, 2020.

sample sizes, which differ across surveys, were determined based on the pilot data for each survey, as described in Appendix Section B. The three surveys are pre-registered on aspredicted.org #47372, #48205 and #49760, respectively. Table 4 displays the sample size for each treatment group for each survey.

Table 5: Demographics Compared to US Population

	Participants	US Population
Prop. Women	0.53	0.51
Median Age	30.00	39.70
Prop. HS Degree or higher	0.99	0.88
Prop. White	0.75	0.76
Prop. Black	0.10	0.13
Prop. Democrat	0.49	0.31
Prop. Republican	0.16	0.31
Prop. Unemployed	0.18	0.37

Notes: N = 6,558. US demographic information is from the 2019 US Census Bureau data. `https://www.census.gov/quickfacts/fact/table/US`, accessed November 2020. US political affiliation is from Gallup, `https://news.gallup.com/poll/15370/party-affiliation.aspx`, accessed November 2020.

While the participant pool is similar to the US population in terms of gender and race, the survey participants tend to be younger, more educated, and are less likely to be unemployed compared to the general US population, as shown in Table 5. Additionally, a higher proportion of survey participants self identify as Democrats relative to the US population. Randomization checks for covariate balance across treatment groups are reported in Tables B6 and B7 in the Appendix.

The title of our survey stated the product category explicitly, the goal being to recruit participants interested in and familiar with that category. In the aluminum survey, 67% of participants report using deodorant daily, 96% of participants brush their teeth at least once daily, and 36% of participants in the GMO survey report purchasing nutritional shakes in the month prior to the survey.

Additionally, we find that the vast majority of participants pass the verification checks for both the ad and debunking sources and content. For instance, 80%-91% of participants answer the debunking source verification check correctly, and about 90% answer the debunking content verification check correctly. We do not see systematic patterns in which sources have lower or higher pass rates across the surveys. As a robustness check, we conduct analysis on both the entire sample and only those who passed the verification checks. The verification check responses are reported in Tables B8 and B9 in the Appendix.

Table 6 displays the share of selected options in the conjoint questions for which the

Table 6: Choice Share for Products with the Focal Ingredient

	Control Ad, No Debunk	Misinfo Ad, No Debunk	Misinfo Ad, Debunk
Aluminum	0.29	0.28	0.39
Fluoride	0.67	0.56	0.64
GMO	0.31	0.29	0.31

Notes: This table displays the share of the selected options containing the focal ingredient among each product category and treatment condition. We group all the debunking sources into one for ease of interpretation.

product contains the focal ingredient across various treatment conditions. First, in the control condition (Table 6, Col. 1), 29% of the selected options contained aluminum, 67% contained fluoride, and 31% contained GMOs. This suggests that there are pre-existing preferences towards or against these ingredients that vary by products. In general, consumers avoid aluminum in deodorant and GMOs in food, but prefer fluoride in toothpastes.

Second, we find that the share of options with the focal ingredient is lower after individuals are exposed to ads containing misinformation (Table 6, Col.2), indicating that misinformation regarding the focal ingredient reduces preferences for the ingredient. Among the group exposed to debunking after misinformation (Table 6, Col.3), the choice share increases relative to Col.2, suggesting that debunking after exposure to misinformation may be effective at reducing the damage created by misinformation. In the following section, we formally estimate the treatment effects, controlling for brand and price effects.

6 Results

Recall that the decision process in the experiment is as follows. After exposure to an ad (control or treated) and a debunking message from a randomly-chosen source ($s \in \{control, competitor, media, regulator\}$), each individual i is presented with 10 sets of product profiles in a sequence. In each set J, the individual compares across 3 random product profiles and the "none" option, and chooses the one that gives her the highest utility in that set. Formally, the probability of individual i choosing product profile j from set J is:

$$Pr(j)_i = \frac{e^{v_{ij}}}{\sum_{k \in J} e^{v_{ik}}} \tag{11}$$

in which the utility from product profile j conditional on individual i's ad and debunking exposure is specified as a series of interaction terms between the ingredient dummy, x_j, and dummies of exposure to control vs. treated ad (I_i^C, I_i^T) and dummies of debunking messages from a given source s ($I_i^s, s \in \{control, competitor, media, regulator\}$):

$$u_{ij} = v_{ij} + \epsilon_{ij} = \sum_s \beta_1^s x_j I_i^C I_i^s + \sum_s \beta_2^s x_j I_i^T I_i^s + \alpha Price_j + Z_j\gamma + \epsilon_{ij} \tag{12}$$

We control for preferences for brands and other balancing attributes via brand fixed effects and balancing attribute fixed effects in Z_j. ϵ is assumed to be *iid* and has an extreme value type 1 distribution. β_1^s captures the average preference for the debated ingredient (i.e., aluminum for deodorant, fluoride for toothpaste, and GMOs for nutritional shakes) under debunking source s across participants in the *control* ad condition, while β_2^s captures the average preference for the debated ingredient in the *treatment* ad condition under debunking source s. We normalize the utility of the outside option, which is selected when the participant selects "None of the above", to 0. We first present the results from the ingredient surveys, and present the results from the belief survey in Section 6.1.

Table 7 reports the estimation results for each product category. We group the two versions of treatment ad for deodorant for ease of interpretation, because they give substantively the same conclusions. Results with all three ad versions are reported in Table B11 in the Appendix. In the sections below, we describe the effects in detail, focusing on the effect of the treatments on the willingness-to-pay (WTP) of the focal ingredient, derived as the ingredient coefficient divided by the absolute value of the price coefficient. We do so because the WTP (as opposed to the ingredient coefficient) enables comparison of effect sizes across products.

Baseline preferences

Our empirical results show that baseline preferences (Table 7, row 1) vary significantly across the three product categories with respondents exhibiting a positive WTP for fluoride in toothpastes and a negative WTP for GMOs in shakes and aluminum in deodorants. The average WTP for aluminum, fluoride and GMO are -\$1.93 (Table 7 column 1, -0.819/0.425), +\$3.27 (Table 7 column 2, 1.426/0.436) and -\$3.60 (Table 7 column 3, -0.614/0.171), respectively. In other words, prior to the experimental manipulation, participants are on average averse to aluminum and GMO; they are willing to pay \$1.93 more for a 2.7oz deodorant without aluminum, and \$3.60 more for a standard case of a dozen GMO-free shakes. For fluoride however, participants are willing to pay an additional \$3.27 for a standard 4oz tube containing fluoride. Figure 9a plots these baseline WTPs across all three ingredients.[16]

[16] Our WTP estimates are comparable to the real-world price differences: a 2.6oz Dove aluminum-free deodorant is priced \$2.60 higher than the version with aluminum, a 12 pack of Ensure protein shakes without GMOs is priced \$6.06 higher than the version with GMOs, and a 4oz. Tom's of Maine toothpaste with fluoride is priced \$1.28 higher than the version without fluoride. As of November 2020 on Amazon.com, a 2.6oz aluminum-free Dove deodorant is \$7.49, and a 2.6oz deodorant with aluminum is \$4.89; the Ensure

Effect of misinformation

Misinformation can alter consumers' preferences in the following ways. If the presented information (e.g., "ingredient is harmful") is different from what she already believes (e.g., "ingredient is *not* harmful"), the consumer will reduce her WTP for the ingredient if she updates in a standard Bayesian fashion. It can also lead to no change in her WTP if she chooses to ignore the conflicting information, a prediction consistent with confirmation bias. If the presented information (e.g., "ingredient is harmful") is consistent with what she already believes, it can increase aversion to the ingredient or lead to no change if there is a ceiling effect.

Figure 9b (Col 1) shows that, for aluminum and GMOs, there is no significant change in WTP after exposure to misinformation. However, the ad containing misinformation causes a statistically significant $1.37 (42%) decrease in the WTP for fluoride. An explanation for why misinformation decreases WTP for fluoride, but not for aluminum and GMOs, may be due to the differences in the baseline preferences across the ingredients: consumers are ex-ante negatively inclined towards aluminum and GMO, but have a strong ex-ante preference for fluoride (Figure 9a). Therefore, an absence of response to misinformation does not necessarily imply that consumers are immune to misleading information: it might merely mean that consumers have strong prior misinformed beliefs, and the additional exposure in this experiment does not shift these beliefs. Therefore, although an additional exposure to misinformation may not have an effect, debunking such misinformation can still play an important role.

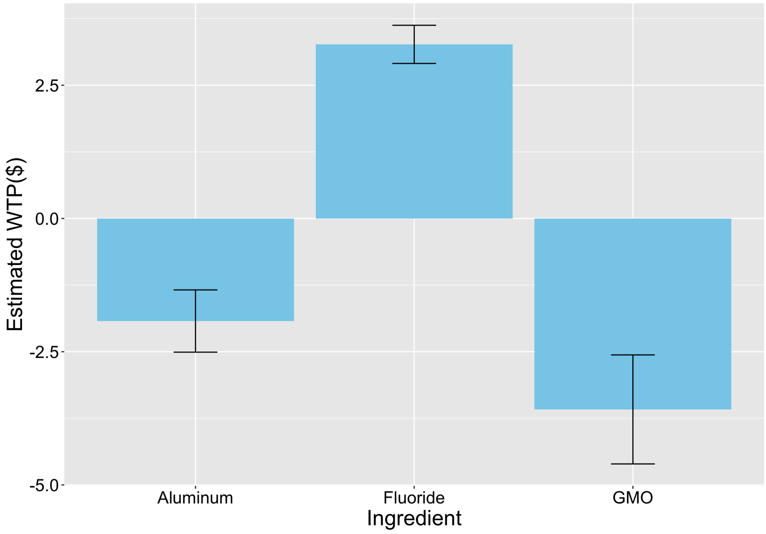
Effect of debunking after experimental exposure to misinformation

Column 2 in Figure 9b displays the net effect of misinformation and debunking. This column allows us to answer whether debunking is able to "undo" the impact of an additional dose of misinformation on WTP. We find that debunking almost entirely reverts consumers to their baseline preference for fluoride: the negative WTP for fluoride under misinformation (Col 1 in Figure 9b) reverts to zero after debunking (Col 2 in Figure 9b). The net effect for aluminum is significantly positive for aluminum despite no effect of misinformation, indicating that debunking is able to impact pre-existing preferences. Overall, debunking is able to revert the negative impact of an additional exposure to misinformation for fluoride and increase consumers' WTP for aluminum potentially correcting pre-existing misinformed beliefs, a possibility we further explore in Section 6.1.

shake without GMO is $0.25/fl. oz, and the Ensure shake with GMOs is $0.20/fl. oz; the Tom's of Maine peppermint toothpaste without fluoride is $0.72/oz, and the one with fluoride is $1.04/oz.

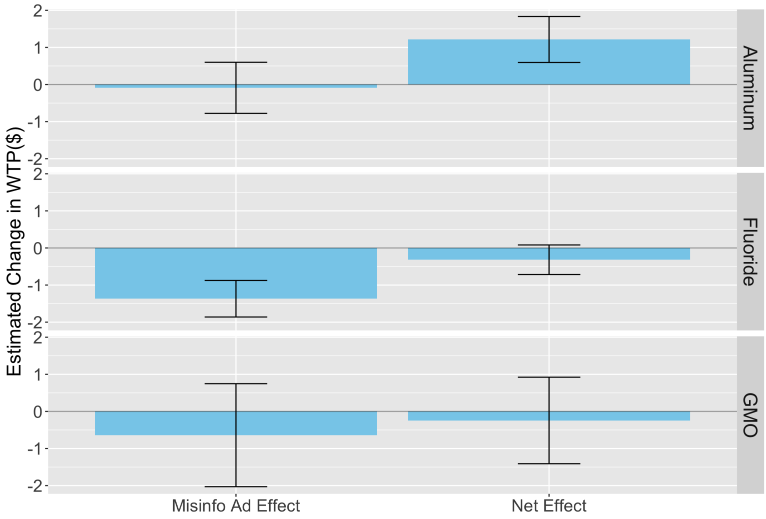
Figure 9: Estimated Ad Effects and Debunking Effects on Willingness-to-Pay.

(a) Baseline WTP



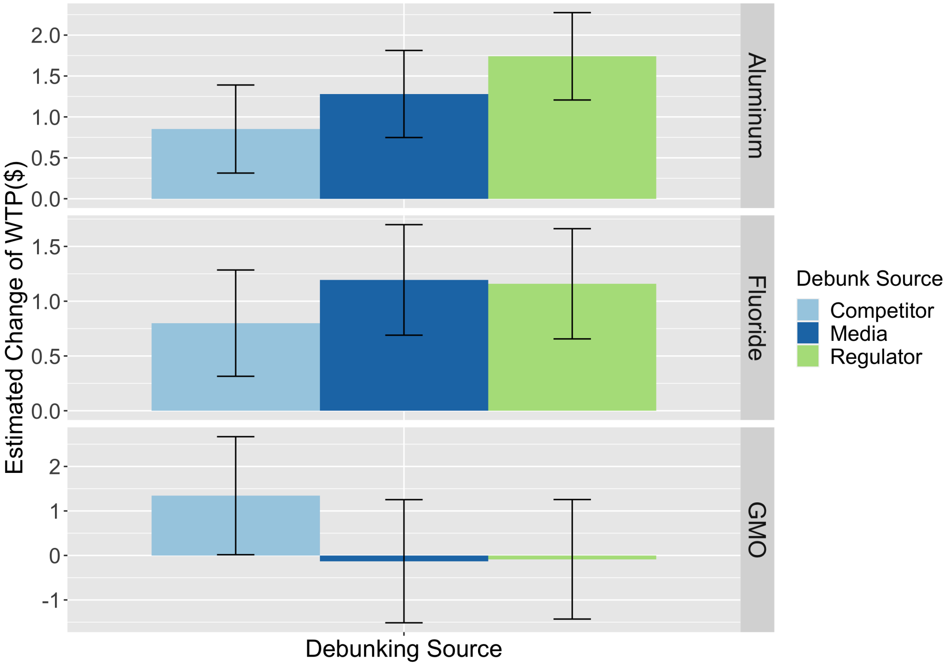
Notes: This figure reports the mean and 95% CI of the willingness–to-pay estimates from the focal ingredients: aluminum (2.7oz deodorants), fluoride (4oz toothpastes), GMOs (a pack of 12 nutritional shakes). Estimates are obtained from individuals in the "Control Ad + No Debunk" condition, with price, brands and other attributes controlled for.

(b) Change in WTP



Notes: This figure displays the average change in the WTP for different treatment conditions. "Debunk Effect" is the effect of debunking on WTP for participants who did not see an ad with misinformation ("Control Ad + Competitor/Media/Regulator Debunking" relative to "Control Ad + No Debunk"). "Misinfo Ad Effect" is the effect of seeing misinformation (comparing participants in the "Misinformation Ad + No Debunk" relative to the "Control Ad + No Debunk"). "Net Effect" is the effect of seeing an ad with misinformation followed by debunking for participants ("Misinformation Ad + Competitor/Media/Regulator Debunking" relative to the "Control Ad + No Debunk"). Error bars represent the 95% confidence intervals.

Figure 10: Estimated Debunking Effects on Willingness-to-Pay by Source and Ingredient.



Notes: This figure displays the average change in the WTP from different debunking sources for participants who saw an ad with misinformation. The comparison baseline is “Misinformation Ad + No Debunk”. Error bars represent the 95% confidence intervals.

Debunking effectiveness by source

Next, we compare the effectiveness of debunking across the sources: competitor firm, media, and regulator. We find overall that regulator debunking causes the largest increase in WTP, followed by media debunking and then by competitor debunking (Figure 10), although these differences across sources are not statistically significant. The only debunking source that is effective at increasing WTP for GMOs is the competitor firm. A potential rationale for the effectiveness of only competitors for GMOs is the high levels of debunking efforts by the media and regulators already in the market. Competitor debunking may be more novel relative to the others, and therefore, more effective.[17]

[17] Google Trends data shows that keyword searches for GMOs in media and from regulators is generally greater than the number of keyword searches in aluminum and fluoride from media and regulators.

Table 7: Main Results for Demand Estimates

	(1)	(2)	(3)
	Aluminum	Fluoride	GMOs
$ingredient * controlad$	-0.819***	1.426***	-0.614***
	(0.126)	(0.0792)	(0.0893)
$ingredient * controlad * competitordebunk$	0.408**	-0.00417	0.0197
	(0.176)	(0.109)	(0.118)
$ingredient * controlad * mediadebunk$	0.312*	0.103	0.0569
	(0.178)	(0.110)	(0.121)
$ingredient * controlad * regulatordebunk$	0.582***	0.0327	0.0554
	(0.166)	(0.109)	(0.122)
$ingredient * misinfoad$	-0.856***	0.829***	-0.723***
	(0.0810)	(0.0806)	(0.0823)
$ingredient * misinfoad * competitordebunk$	0.362***	0.349***	0.230**
	(0.116)	(0.108)	(0.116)
$ingredient * misinfoad * mediadebunk$	0.545***	0.521***	-0.0223
	(0.114)	(0.112)	(0.121)
$ingredient * misinfoad * regulatordebunk$	0.741***	0.506***	-0.0148
	(0.114)	(0.112)	(0.117)
$price$	-0.425***	-0.436***	-0.171***
	(0.0153)	(0.0121)	(0.00622)
Control dummies	balancing attributes, brands		
N	1,797	3,202	1,559

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered by individuals.

Notes: Balancing attributes are: "scented" for deodorant, "whitening" for toothpaste, and "vanilla/chocolate" for nutritional shakes. Brands included in the studies are: {Kopari, Dove, Speed Stick} for deodorant, {Colgate, Crest, Tom's of Maine, Risewell} for toothpastes, and {Soylent, Orgain, Ensure} for nutritional shakes. We group the weak and strong versions of misinformed ads in deodorant study together. The ungrouped regression estimates for the aluminum survey are reported in Table B11 in the Appendix. Table B10 in the Appendix displays the results for only those who pass the verification checks.

6.1 Heterogeneity by Ex-Ante Beliefs

Here, we discuss results from the beliefs survey, designed to explicitly understand whether debunking has the most impact on consumers with the most misinformed beliefs. As highlighted in Section 4.4, respondents' beliefs on the harmfulness of the ingredient were elicited prior to being exposed to the treatment. After treatment, along with the choice data which provides us with the willingess-to-pay estimates, we again collect respondents' beliefs as a metric of consumers' posterior beliefs.[18]

We implement this survey for the deodorant product category because a) consumers are more likely to have misinformed beliefs as seen by the negative WTP for the ingredient in Figure 9a, as opposed to toothpaste where most people are favorable to fluoride, and b) we are able to get more precise and economically meaningful responses to the debunking message as can be seen in Figure 9b, as opposed to GMOs. Because our goal is to detect significant differences across belief groups, we ran a pilot survey to estimate the required sample size, resulting in N=240 respondents per belief group. With 2 advertisements (control ad vs. misinformation ad) and 2 debunking messages (control vs. regulator), and 5 belief groups, this leads to a total sample size of $2 \times 2 \times 5 \times 240 = 4800$, a fairly large number. We therefore restrict attention to just one product category, and one debunking source for budgetary reasons. We pick the regulator as the debunking source because it has the largest impact on consumers in our previous surveys.

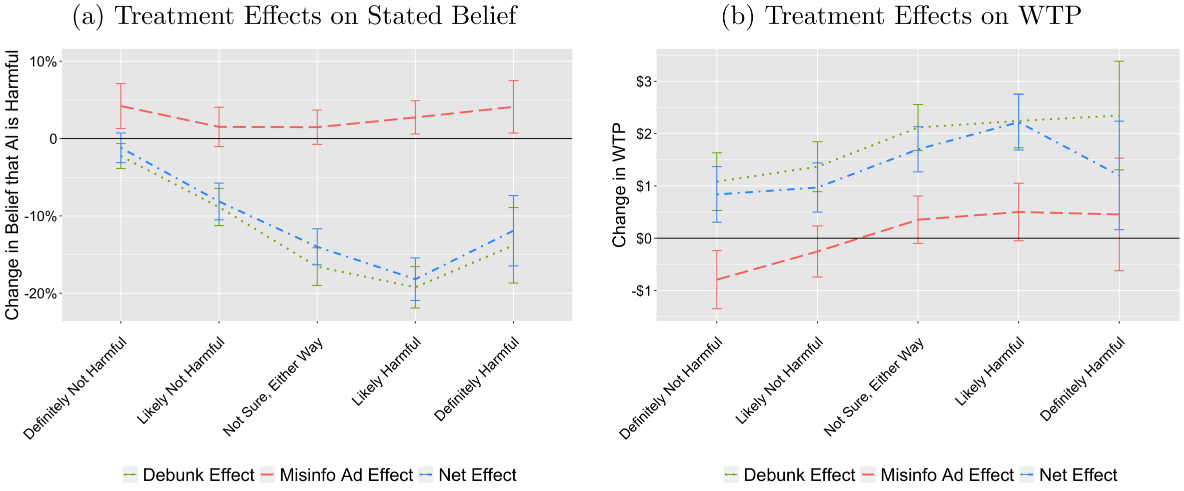
This survey was implemented in March 2021, with a sample size of $N = 4,758$, pre-registered on aspredicted.org #62331. Participants that have taken one of the previous three surveys are excluded from taking this survey to avoid familiarity with the experiment design. Table B12 displays the characteristics of the participants in this survey.

Figure 11 reports the estimated treatment effects on stated beliefs and willingness-to-pay, under exposure to the misinformation ad only, to debunking only, and to both the misformation ad and debunking (net effect). To obtain the treatment effects on stated beliefs for each belief group (Figure 11a), we compare the within-person differences between the ex-post and the ex-ante stated beliefs about the focal ingredient across the three treatment arms and the control arm.

First, we find that misinformation impacts preferences by creating misbeliefs - consumers who had prior beliefs that aluminum is not harmful become more likely to believe that aluminum is harmful and reduce their willingness-to-pay after being exposed to misinformation. We also find some evidence that misinformation strengthens misinformed beliefs - participants with priors that aluminum is harmful are more likely to state they believe aluminum is harmful after treatment. However, this effect did not translate into a change in WTP for these participants. Consistent with the first aluminum survey, the effect of misinformation (red dashed line) is significantly smaller than the effect of the debunking message (green dotted line), leading to a robust net effect (blue dot-dashed line) where all consumers are now less likely to believe the ingredient is harmful. Overall, an additional dose of misinformation

[18] Because of repeated measure of the same metric (beliefs), reversion to the mean is of concern. As an example, those who state they believe the ingredient is "Definitely Not Harmful" are more likely to state one of the other choices making it seem they responded to treatment while in fact it is a statistical artifact of repeated measurements. We use the control group to control for reversion to the mean.

Figure 11: Estimated Change in Stated Belief and WTP


(a) Treatment Effects on Stated Belief

(b) Treatment Effects on WTP

Notes: The figures plot the estimated treatment effect and 95% CI on stated beliefs (left panel) and WTP (right panel). The x-axis is the individual's ex-ante stated belief that aluminum is harmful. The red dashed line represents comparison across the control group and the misinfo-ad group. The green dotted line represents comparison across the control group and the debunking group. The blue dot-dashed line represents comparison across the control group and the group that is exposed to both misinfo ad and debunking.

does not make debunking significantly less effective (i.e., green and blue lines overlap).

Second, debunking is most effective for consumers with priors "Likely Harmful" as can be seen by the steep decline in stated beliefs for this group. Comparing Figure 11a to Figure 2c in the Theory section provides evidence consistent with rational Bayesian updating. On the contrary, confirmation bias would predict that consumers with priors on the left hand side of the x-axis ("Likely Not Harmful") would update the most because they view the debunking message as consistent with their existing beliefs. Overall, we find good news for policy makers: the evidence is supportive of the ideal outcome that debunking has a larger impact on consumers with more misinformed beliefs.

We confirm these patterns using the willingess-to-pay estimates constructed directly from the conjoint choice data (as done in Section 6). The willingness-to-pay patterns (Figure 11b) are consistent with the stated beliefs. Specifically, the impact from debunking across the belief groups (green dotted line) significantly outweighs the impact from misleading ad (red dashed line). Combining the net effect from misleading ad and debunking, we find a significant increase in willingness-to-pay across the board (blue dot-dashed line). Similar as before, we find the largest net effect occurring with people who think the ingredient is "likely harmful": everything else constant, they are now willing to pay $2.20 more for a 2.7oz deodorant with aluminum. This pattern in willingness-to-pay also provides evidence that the effects in the stated beliefs are not driven by ceiling effects (i.e., those who state they consider the ingredient to be “Definitely Not Harmful” have less room to move than those who state “Likely Harmful”), as willingess-to-pay estimates are unbounded.

6.2 Incentives to spread misinformation

In this section, we investigate the incentives for new products to make misleading claims at market entry. We also investigate whether incumbents have incentives to debunk such misleading claims or instead simply introduce a new product without the debated ingredient, succumbing to the presence of such misleading claims. We do so by simulating the market share of the new brand and the incumbents, with and without misinformation / debunking, at the time of: 1) new brand market entry; and 2) introduction of a revised product without the debated ingredient by the incumbent.

We focus on the toothpaste category because, as shown in Figure 9b above, misinformation has the largest impact on consumers' WTP for the active ingredient fluoride. To simulate market shares, we use the estimates from the fluoride survey in Column 2 of Table 7. We consider three incumbents - Colgate, Crest and Tom's of Maine - using their Amazon best-seller configurations (e.g., whitening toothpaste with fluoride) and price, and consider Risewell as the new entrant.

Our market share estimates are similar to actual US market shares for Colgate and Crest, which were 33% and 34.7%, respectively (WSJ.com, 2018). In 2020, fewer than 5% of households reported using a toothpaste brand in the luxury oral-care segment, where Risewell and dozens of other brands belong.[19] Nevertheless, these beauty-focused oral-care brands are gaining a foothold in the US market, exploiting controversies around ingredients through social media marketing and upscale branding (Glossy.com, 2020).

First, we find that misinformation about the focal ingredient helps the new entrant. Specifically, making misleading claims in ads about fluoride toxicity increases the market share of Risewell by almost two-fold (Table 8, 0.27% vs. 0.16%). Such misinformation hurts all incumbents with fluoride, leading to decreased market shares as shown in Table 8, Column “Entry with Misinfo”.

Second, we find that although debunking by one major incumbent helps partially recover its market share, complying with the misinformation, and thus misinformed beliefs,improves its share even more. Specifically, Colgate increases its share by 1.29% after debunking the misinformation (Table 9, 33.11% vs. 31.82%). Other incumbents also gain from Colgate's debunking efforts suggesting positive spillovers of the debunking message. Moreover, Colgate's increase in market share is much lower than the gain Colgate achieves when it introduces a

[19] All the luxury oral-care brands (including Theodent, Risewell, Marvis, Aesop, Native, Hello, Quip, Bite, Boka, etc.), have less than 5% share of self-reported usage in total (Statista, 2020).

Table 8: Simulated Market Share at Risewell's Entry

Brand	Price per Tube	Simulated Market Share		
		Before Entry	Entry with Misinfo	Entry without Misinfo
Colgate	$3.52	33.89%	31.82%	33.83%
Crest	$3.32	38.41%	36.08%	38.35%
Tom's of Maine	$4.33	20.13%	18.90%	20.10%
Risewell	$12	0	0.27%	0.16%

Notes: The table reports the simulated market share for Risewell and for three competitor brands. All products offer extra benefit of whitening. Only Risewell comes without fluoride. The price per tube is calculated based on the Amazon best-seller/Amazon's choice search result for each competitor brand, and from Risewell's website. Market share is simulated using estimates from Table 7.

fluoride-free toothpaste instead of debunking: the total market share for Colgate from products with and without fluoride reaches 40.14%, more than its market share before Risewell's entry (33.89%).[20] More importantly, Colgate becomes the market leader once it complies with the misinformation. In this case, the other incumbents, Crest and Tom's of Maine, are significantly hurt (with their shares further dropping by 4.4% and 2.3% respectively).

Table 9: Simulated Market Share after Risewell's Entry with Misinformation

Brand	Price per Tube	Simulated Market Share after Entry with Misinformation		
		No Debunking	Colgate Debunking	Colgate Introducing Fl-free
Colgate	$3.52	31.82%	33.11%	40.14%
Crest	$3.32	36.08%	37.54%	31.68%
Tom's of Maine	$4.33	18.90%	19.67%	16.60%
Risewell	$12	0.27%	0.20%	0.24%

Notes: The table reports the simulated market share for Risewell and for 3 competitor brands after Risewell's entry with misinformation. All products offer extra benefit of whitening. Col.4 reports the simulated market share when Colgate debunks the misinformation. Col.5 reports the simulated total market share when Colgate offers both fluoride and fluoride-free toothpastes without debunking. Crest and Tom's of Maine only offer toothpaste with fluoride. The price per tube is calculated based on the Amazon best-seller/Amazon's choice search result for each competitor brand, and from Risewell website. Market Share is simulated using estimates from Table 7.

There are two main takeaways from the market share simulation exercise. First, new products have a clear incentive to spread misinformation in order to enter the market more successfully - misinformation serves as a product differentiation tool. Second, incumbent brands lack the incentive to debunk because 1) debunking helps competitors more than their own products, 2) complying with the misinformation improves competitiveness of the incumbent brand, and 3) even when doing nothing, some incumbent brands could benefit

[20] In the simulation, Crest and Tom's of Maine only maintain one product line with fluoride.

from the misinformation.[21] We find the same effect for deodorants (aluminum) and nutrition shakes (GMO), where the brands are better served introducing new product lines without the debated ingredient rather than debunk. Admittedly, without the cost of production, one cannot comment on the relative profitability of introducing a new product line that exploits misbeliefs. Nevertheless, our simulation provides an explanation why so many brands are willing to spread misinformation but few appear ready to debunk.

7 Conclusion

This paper investigates the extent that debunking via corrective messaging can revert the effects of misinformation, and the heterogeneous impacts of misinformation and debunking based on the debunking source and prior beliefs. Through an incentive compatible survey experiment, we measure these impacts on consumers' revealed preferences and stated beliefs. We find that misinformation can influence consumers' willingness to pay, and that debunking provides an effective strategy to revert consumers' beliefs, and in some cases, can even influence consumers' pre-existing beliefs.

While debunking is shown to be effective on average, the hetereogenous impacts of debunking is an important consideration in policy evaluation. From a regulator's perpsective, debunking ideally should change actions resulting from misinformed beliefs (consistent with rational Bayesian updating) rather than reinforcing those with correct beliefs (consistent with confirmation bias). We indeed find this to be true. Directly eliciting prior beliefs from survey participants, we show that debunking is most effective for those who had misinformed beliefs, an encouraging finding for policy makers.

Another important dimension of heterogeneity we consider is the source of the debunking message. Debunking messages from regulators, competitors, and media are effective for aluminum and fluoride, while for GMOs, only the debunking message from the competing firm is effective. The finding that competitor debunking is effective is promising news to regulators who may not have resources to debunk or remove every source of misinformation in the marketplace. However, debunking may not be the incumbent's most profitable strategy - simulations show that introducing a product that conforms to the misinformation leads to a greater increase in market share than debunking. Since competitors may lack the incentive to debunk, existing laws such as the Lanham Act and accreditation sites such as the Better Business Bureau might help bring questionable claims to the attention of regulators.

Methodologically, this paper develops an incentive compatible survey to measure con-

[21] For example, Tom's of Maine sells fluoride-free toothpastes. Presence of misinformation benefits such product lines. A simulation with fl-free Tom's of Maine products confirms this effect.

sumers' responses to misinformation and sources of debunking. This method is straightforward to use and can be applied across a wide variety of categories, especially in settings where field experiments are not feasible. Because consumer beliefs are constantly evolving, future work would benefit from investigating the strength of debunking over time and across categories.

While this paper shows that debunking is effective in a controlled setting where consumers are paying attention to the source and content of the tweets, we are not able to comment on whether these results can hold in a setting where consumers may pay less attention, or can seek out information in a biased manner. Understanding debunking effectiveness when individuals selectively pay attention is an interesting avenue for future research.

References

AFS (2021). Available at `https://americanfluoridationsociety.org/debunking-anti-claims/myths/` (Accessed June 2021).

Armstrong, G. M., M. N. Gurol, and F. A. Russ (1983). A longitudinal evaluation of the listerine corrective advertising campaign. *Journal of Public Policy & Marketing 2*(1), 16–28.

Avaaz (2020). Available at `https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation` (Accessed February 2021).

Avery, R. J., J. H. Cawley, M. Eisenberg, and J. Cantor (2013). Raising red flags: The change in deceptive advertising of weight loss products after the federal trade commission's 2003 red flag initiative. *Journal of Public Policy & Marketing 32*(1), 129–139.

Bago, B., D. G. Rand, and G. Pennycook (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general*.

Barrera, O., S. Guriev, E. Henry, and E. Zhuravskaya (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics 182*, 104123.

Basch, C. H., N. Milano, and G. C. Hillyer (2019). An assessment of fluoride related posts on instagram. *Health promotion perspectives 9*(1), 85.

Bernhardt, K. L., T. C. Kinnear, and M. B. Mazis (1986). A field study of corrective advertising effectiveness. *Journal of Public Policy & Marketing 5*(1), 146–162.

Bollinger, B., P. Leslie, and A. Sorensen (2011). Calorie posting in chain restaurants. *American Economic Journal: Economic Policy 3*(1), 91–128.

Bursztyn, L., A. Rao, C. P. Roth, and D. H. Yanagizawa-Drott (2020). Misinformation during a pandemic. Technical report, National Bureau of Economic Research.

CDC (2020). Community water fluoridation. Available at `https://www.cdc.gov/fluoridation/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Ffluoridation%2Findex.htm`.

Charness, G. and D. Levin (2005). When optimal choices feel wrong: A laboratory study of bayesian updating, complexity, and affect. *American Economic Review 95*(4), 1300–1309.

Chen, K., A. Chen, J. Zhang, J. Meng, and C. Shen (2020). Conspiracy and debunking narratives about covid-19 origins on chinese social media: How it started and who is to blame. *Harvard Kennedy School Misinformation Review*.

Ching, A. T., T. Hossain, S. S. Tehrani, and C. Y. Zhao (2021). How do people update beliefs? evidence from the laboratory. *working paper*.

Chiou, L. and C. Tucker (2018). Fake news and advertising on social media: A study of the anti-vaccination movement. Technical report, National Bureau of Economic Research.

Coutts, A. (2019). Good news and bad news are still news: Experimental evidence on belief updating. *Experimental Economics 22*(2), 369–395.

Delavande, A. (2008). Pill, patch, or shot? subjective expectations and birth control choice. *International Economic Review 49*(3), 999–1042.

Dhar, T. and K. Baylis (2011). Fast-food consumption and the ban on advertising targeting children: the quebec experience. *Journal of Marketing Research 48*(5), 799–813.

Dias, N., G. Pennycook, and D. G. Rand (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review 1*(1).

Ding, M., R. Grewal, and J. Liechty (2005). Incentive-aligned conjoint analysis. *Journal of marketing research 42*(1), 67–82.

Dyer, R. F. and P. G. Kuehl (1974). The "corrective advertising" remedy of the ftc: An experimental evaluation: An empirical study of the effectiveness of corrective advertising. *Journal of Marketing 38*(1), 48–54.

Dyer, R. F. and P. G. Kuehl (1978). A longitudinal study of corrective advertising. *Journal of Marketing Research 15*(1), 39–48.

El-Gamal, M. A. and D. M. Grether (1995). Are people bayesian? uncovering behavioral strategies. *Journal of the American statistical Association 90*(432), 1137–1145.

Feldmann, M. P., M. L. Morris, and D. Hoisington (2000). Genetically modified organisms: Why all the controversy? *Choices 15*(316-2016-6999).

Fernbach, P. M., N. Light, S. E. Scott, Y. Inbar, and P. Rozin (2019). Extreme opponents of genetically modified foods know the least but think they know the most. *Nature Human Behaviour 3*(3), 251–256.

Flynn, D., B. Nyhan, and J. Reifler (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology 38*, 127–150.

FTC (2019). Available at `https://www.ftc.gov/news-events/press-releases/2019/12/ftc-obtains-record-191-million-settlement-university-phoenix` (Accessed September 2020).

FTC (2020). Available at `https://www.ftc.gov/enforcement/cases-proceedings` (Accessed September 2020).

Funk, C., B. Kennedy, and M. Hefferon (2018). Public perspectives on food risks. *Pew Research Center*.

Glossy.com (2020). Available at `https://www.glossy.co/beauty/oral-care-is-a-pandemic-bright-spot-for-beauty/` (Accessed June 2021).

Green, P. E. and V. R. Rao (1971). Conjoint measurement-for quantifying judgmental data. *Journal of Marketing research 8*(3), 355–363.

Green, P. E. and V. Srinivasan (1978). Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research 5*(2), 103–123.

Green, P. E. and V. Srinivasan (1990). Conjoint analysis in marketing: new developments with implications for research and practice. *Journal of marketing 54*(4), 3–19.

Grether, D. M. (1980). Bayes rule as a descriptive model: The representativeness heuristic. *The Quarterly journal of economics 95*(3), 537–557.

Guess, A., J. Nagler, and J. Tucker (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances 5*(1), eaau4586.

Guess, A., B. Nyhan, and J. Reifler (2018). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council 9*(3), 4.

Guess, A. M., B. Nyhan, and J. Reifler (2020). Exposure to untrustworthy websites in the 2016 us election. *Nature human behaviour 4*(5), 472–480.

He, S., B. Hollenbeck, and D. Proserpio (2020). The market for fake reviews. *Available at SSRN*.

Hobin, E., B. Bollinger, J. Sacco, E. Liebman, L. Vanderlee, F. Zuo, L. Rosella, M. L'abbe, H. Manson, and D. Hammond (2017). Consumers' response to an on-shelf nutrition labelling system in supermarkets: evidence to inform policy and practice. *The Milbank Quarterly 95*(3), 494–534.

Holt, C. A. and A. M. Smith (2009). An update on bayesian updating. *Journal of Economic Behavior & Organization 69*(2), 125–134.

Hovland, C. I. and W. Weiss (1951). The influence of source credibility on communication effectiveness. *Public opinion quarterly 15*(4), 635–650.

Ippolito, P. M. and A. D. Mathios (1990). The regulation of science-based claims in advertising. *Journal of Consumer Policy 13*(4), 413–445.

Ippolito, P. M. and A. D. Mathios (1995). Information and advertising: The case of fat consumption in the united states. *The American Economic Review 85*(2), 91–95.

Jindal, P. and A. Aribarg (2021). The importance of price beliefs in consumer search. *Journal of Marketing Research 58*(2), 321–342.

Kahan, D. M. (2008). Cultural cognition as a conception of the cultural theory of risk. *HANDBOOK OF RISK THEORY, S. Roeser, ed., Forthcoming*, 08–20.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Kong, X. and A. Rao (2021). Do made in usa claims matter? *Marketing Science 0*, 1–35.

Lazer, D. M., M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. (2018). The science of fake news. *Science 359*(6380), 1094–1096.

Lewandowsky, S., U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest 13*(3), 106–131.

Liaukonyte, J., N. A. Streletskaya, H. M. Kaiser, and B. J. Rickard (2013). Consumer response to "contains" and "free of" labeling: Evidence from lab experiments. *Applied Economic Perspectives and Policy 35*(3), 476–507.

Linvill, D. L. and P. L. Warren (2020). Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, 1–21.

Lorenz-Spreen, P., M. Geers, T. Pachur, R. Hertwig, S. Lewandowsky, and S. M. Herzog (2020). A simple self-reflection intervention boosts the detection of targeted advertising.

Luca, M. and G. Zervas (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science 62*(12), 3412–3427.

Manski, C. F. (2004). Measuring expectations. *Econometrica 72*(5), 1329–1376.

Martel, C., M. Mosleh, and D. Rand (2020). You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online.

Mayzlin, D., Y. Dover, and J. Chevalier (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review 104*(8), 2421–55.

Mazis, M. B. and J. E. Adkinson (1976). An experimental evaluation of a proposed corrective advertising remedy. *Journal of Marketing Research 13*(2), 178–183.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology 2*(2), 175–220.

Nyarko, Y. and A. Schotter (2002). An experimental study of belief learning using elicited beliefs. *Econometrica 70*(3), 971–1005.

Olson, J. C. and P. A. Dover (1978). Cognitive effects of deceptive advertising. *Journal of Marketing Research 15*(1), 29–38.

Penn Medicine (2019). Is deodorant harmful for your health? Available at `https://www.pennmedicine.org/updates/blogs/health-and-wellness/2019/june/deodorant`.

Pennycook, G., A. Bear, E. T. Collins, and D. G. Rand (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*.

Pennycook, G., Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand (2021). Shifting attention to accuracy can reduce misinformation online. *Nature 592*(7855), 590–595.

Pennycook, G., J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science 31*(7), 770–780.

Pennycook, G. and D. G. Rand (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences 116*(7), 2521–2526.

Pennycook, G. and D. G. Rand (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition 188*, 39–50.

Pratkanis, A. R., A. G. Greenwald, M. R. Leippe, and M. H. Baumgardner (1988). In search of reliable persuasion effects: Iii. the sleeper effect is dead: Long live the sleeper effect. *Journal of personality and social psychology 54*(2), 203.

Prolific.com (2021). See `https://www.prolific.co`.

Rabin, M. and J. L. Schrag (1999). First impressions matter: A model of confirmatory bias. *The quarterly journal of economics 114*(1), 37–82.

Rao, A. (2020). Deceptive claims using fake news marketing: The impact on consumers. *Available at SSRN 3248770*.

Rao, A. and E. Wang (2017). Demand for "healthy" products: False claims and ftc regulation. *Journal of Marketing Research 54*(6), 968–989.

Schotter, A. and I. Trevino (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ. 6*(1), 103–128.

Schwarz, N. and M. Jalbert (2020). When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. *The psychology of fake news: Accepting, sharing, and correcting misinformation*, 113–139.

Schwarz, N., E. Newman, and W. Leach (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy 2*(1), 85–95.

Scott, S. E. and P. Rozin (2020). Actually, natural is neutral. *Nature Human Behaviour*, 1–2.

Shin, S., S. Misra, and D. Horsky (2012). Disentangling preferences and learning in brand choice models. *Journal of Marketing Research 31*(1), 115–137.

Simonov, A., S. K. Sacher, J.-P. H. Dubé, and S. Biswas (2020). The persuasive effect of fox news: non-compliance with social distancing during the covid-19 pandemic. Technical report, National Bureau of Economic Research.

Skurnik, I., C. Yoon, D. C. Park, and N. Schwarz (2005). How warnings about false claims become recommendations. *Journal of Consumer Research 31*(4), 713–724.

Speedstick (2020). Available at `https://www.speedstick.com/en-us/sweat/aluminum-free-deodorant-is-it-safe` (Accessed October 2020).

Statista (2020). Available at `https://www.statista.com/statistics/278185/us-households-brands-of-toothpaste-used/` (Accessed June 2021).

Thaler, R. H. and C. R. Sunstein (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Ursu, R., Q. Zhang, and T. Erdem (2021). Prior information and consumer search: Evidence from eye-tracking. *Available at SSRN 3717885*.

Vosoughi, S., D. Roy, and S. Aral (2018). The spread of true and false news online. *Science 359*(6380), 1146–1151.

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

Walter, N. and S. T. Murphy (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs 85*(3), 423–441.

WebMD.com (2011). Available at `https://www.webmd.com/breast-cancer/features/antiperspirant-facts-safety` (Accessed Jan 2021).

Wilkie, W. L., D. L. McNeill, and M. B. Mazis (1984). Marketing's "scarlet letter": The theory and practice of corrective advertising. *Journal of Marketing 48*(2), 11–31.

Wilson, A. (2014). Bounded memory and biases in information processing. *Econometrica 82*(6), 2257–2294.

WSJ.com (2018). Available at `https://www.wsj.com/articles/brushers-are-buying-sensodynes-pitch-1534942800` (Accessed June 2021).

Appendix

A Confirmation Bias

Confirmation bias can manifest in one of the following ways: individuals can seek out information that is consistent with their hypothesis, and individuals overweight evidence that confirms their hypothesis and underweight evidence that conflicts with their hypothesis (Nickerson, 1998). Our study context is more in line with the latter. Specifically, we model the weighting of evidence by how trustworthy the individual perceives the source to be; individuals view a source to be less trustworthy the more the source's information differs from their priors.

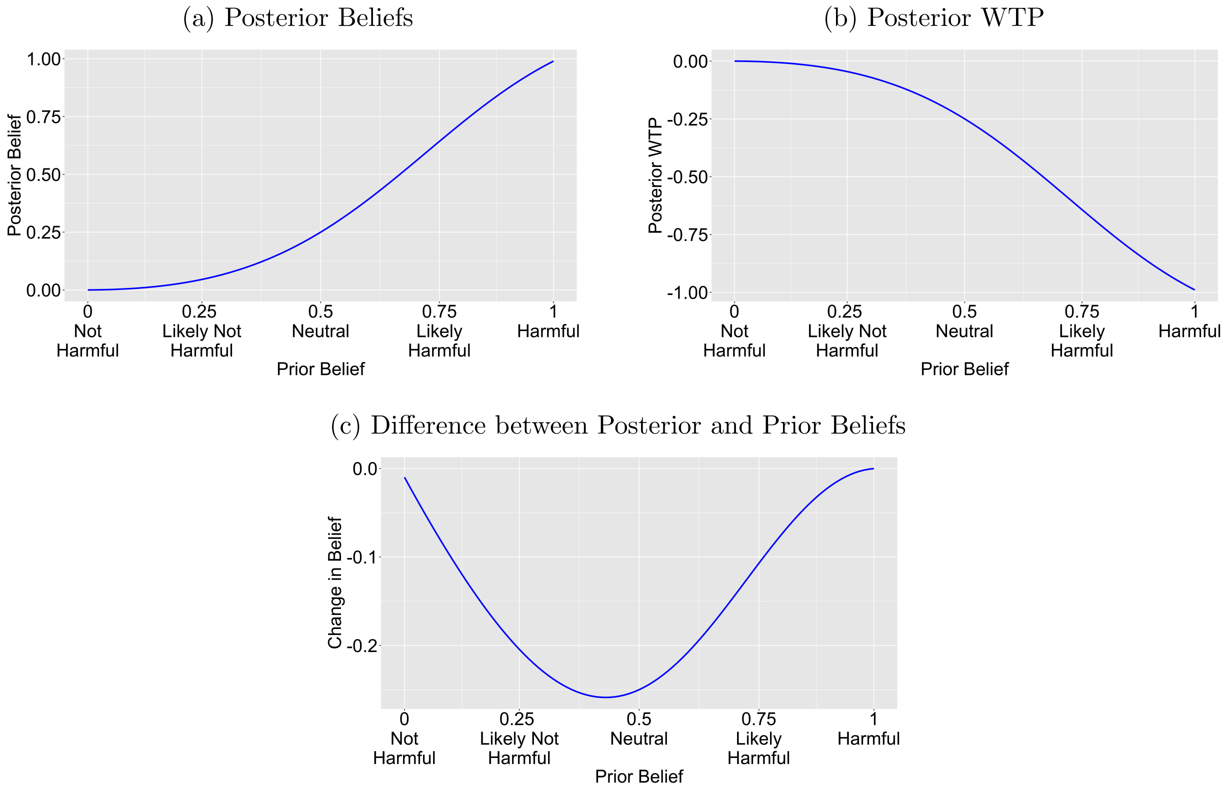
To provide an example of how beliefs are updated under confirmation bias, we apply the following functional form of how a source's trustworthiness is perceived. If an individual's prior belief is that the ingredient is definitely not harmful ($\theta^0 = 0$) and she receives information that the ingredient is not harmful ($d = \texttt{claim not harmful}$), then her perceived trustworthiness of the source, denoted by $\tilde{\pi}$ is 1. However, if her prior is that the ingredient is definitely not harmful and she received information that the ingredient *is* harmful ($d = \texttt{claim harmful}$), then $\tilde{\pi} = 0.5$, or in other words, the source's information is completely uninformative. The more the information conflicts with her prior, the less trustworthy she perceives the source to be, therefore the less weight she puts on the information. Formally, the perceived trustworthiness, $\tilde{\pi}$, is updated to a value in the range $[0.5, 1]$ as a function of the prior and the provided information:

$$\tilde{\pi} = \begin{cases} \frac{\theta^0}{2} + 0.5 & d = \texttt{claim harmful} \\ \frac{1-\theta^0}{2} + 0.5 & d = \texttt{claim not harmful} \end{cases} \tag{13}$$

The posterior is updated in the same procedure in Table 1 with π being replaced with $\tilde{\pi}$. While there exist other models of confirmation bias[22], we chose to specify the biased weighting of information through the trustworthiness of the source in order to make both the confirmation bias and standard Bayesian updating frameworks in this paper comparable. Figure A1 plots the posterior beliefs, posterior WTP, and change in beliefs when individuals update in complete accordance with confirmation bias. Figure A1c, which displays the change in beliefs relative to the prior, shows that individuals who ex-ante believed that the ingredient

[22] Examples of other ways that confirmation bias has been specified is by assigning a non-zero probability that an individual misreads conflicting information to be supporting his prior (Rabin and Schrag, 1999), and through bounded memory, in which individuals can recall a finite number of pieces of information when updating beliefs, and are more likely to recall information that supports their priors (Wilson, 2014).

Figure A1: Confirmation Bias Posterior Beliefs and WTP


(a) Posterior Beliefs

(b) Posterior WTP

(c) Difference between Posterior and Prior Beliefs

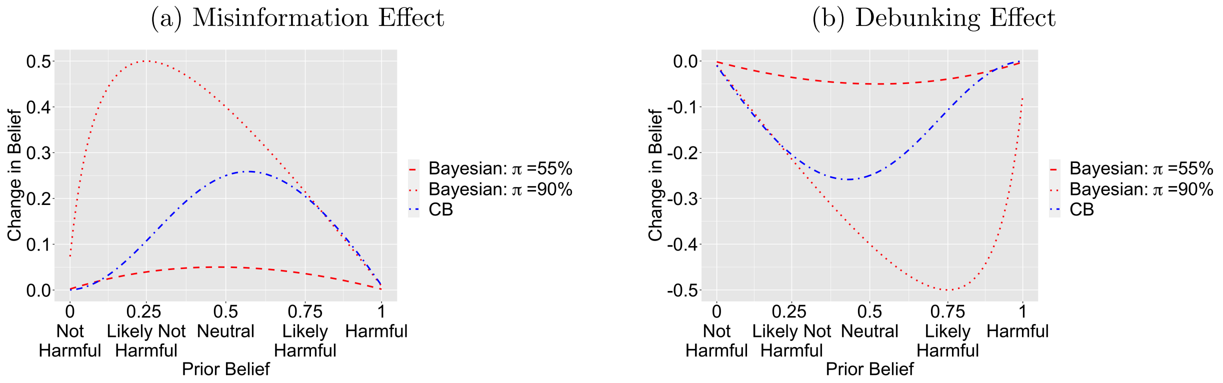
Notes: Figures A1a and A1b displays θ^{post}, an individual's posterior belief that the ingredient is harmful, and her corresponding WTP, respectively, for the ingredient after she sees a debunking message. The posterior belief is derived from the equations in Table 1, the individual's perceived trustworthiness in Equation 13, and the WTP is derived from the posterior belief at the following parameter values $\gamma = 1$, $\tau = 0$, and $\delta = -1$.

is not harmful but are uncertain are the most responsive to debunking. A noteworthy feature of Figure A1c is the left-skewed nature of the curve, compared to the right skewed curve of Figure 2c. This left-skewed curve captures the intuition that according to confirmation bias, those whose beliefs are not that opposed to the content of the debunking message are willing to update whereas those whose beliefs are contradictory to the content of the debunking message are less willing to update.

A.1 Comparing Model Predictions on Posterior Beliefs

Figure A2 compares the changes in beliefs under two frameworks. The most notable differences between the two frameworks are which individuals the treatment is the most effective for, based on their priors. Summarizing these differences, the frameworks make the following predictions.

Figure A2: Change in Beliefs for Bayesian Updating vs. Confirmation Bias


(a) Misinformation Effect

(b) Debunking Effect

Notes: Figures A2a and A2b plot the change in beliefs (posterior - prior) relative to the prior when the individual is exposed to misinformation ($d = claimtoxic$) and debunking ($d = claimenottoxic$), respectively. For ease of illustration, the trustworthiness of the source is set to be 0.55 and 0.9 in the posteriors updated via Bayes rule.

Prediction 1 (BU): If individuals fully follow Bayes rule, individuals with priors $\theta^0 > 0.5$ update their posteriors the most in response to debunking, and individuals with priors $\theta^0 < 0.5$ update their priors the most in response to misinformation.
Prediction 2 (CB): If individuals fully follow confirmation bias, individuals with priors $\theta^0 < 0.5$ update their posteriors the most in response to debunking, and individuals with priors $\theta^0 > 0.5$ update their priors the most in response to misinformation.

Prediction 3 (*Trustworthiness*): When the source's trustworthiness is low, Bayes rule and conformation bias lead to similar posteriors.

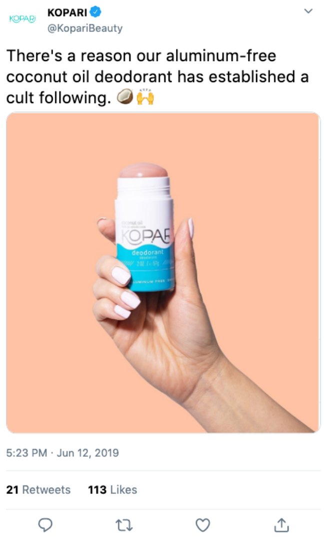
Prediction 3 guides the experimental design: to be able to disentagle the two frameworks empirically, the information must come from a trustworthy source. In the following sections, we use Predictions 1 and 2 to guide the interpretation of our empirical results.

B Power Calculation

Before launching the survey for each product category, we ran a pilot survey with 10 respondents to determine the sample size required to conduct inference. The pilot was deployed with one control ad arm and one treatment ad arm (both under control debunking) and was used to determine the standard deviation and effect size of the ingredient attribute. This process helped determine the required sample size as 75 per treatment arm for aluminum, 200 for fluoride, and 100 for GMOs. In the survey implementation, we doubled the required sample sizes to be conservative. These are documented in the pre-registration links created before the survey at aspredicted.org.[23]

C Additional Figures and Tables

Figure B1: Additional Treatment Ad in Aluminum Survey



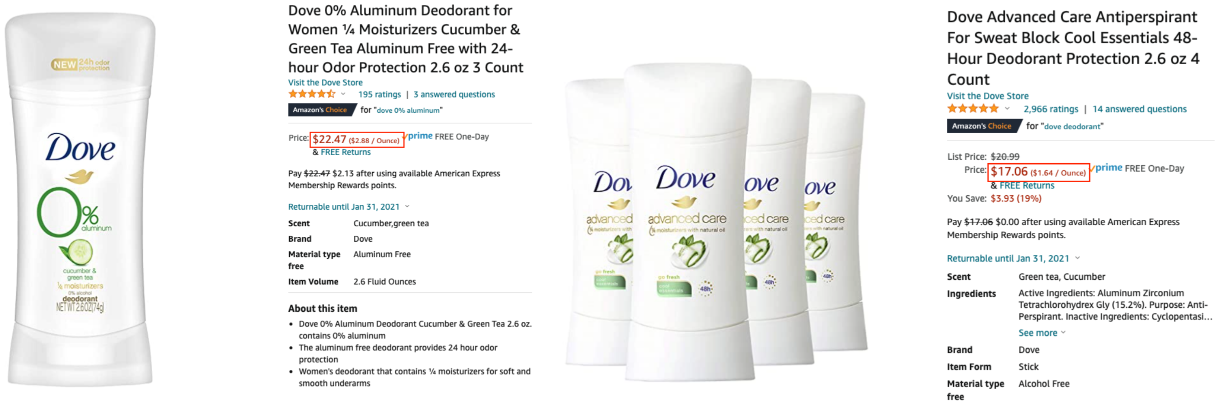
KOPARI
@KopariBeauty

There's a reason our aluminum-free coconut oil deodorant has established a cult following. 🥥🙌

5:23 PM · Jun 12, 2019

21 Retweets **113** Likes

Figure B2: Product Pricing Examples - Deodorant With and Without Aluminum



(a) Aluminum-free

Dove 0% Aluminum Deodorant for Women ¼ Moisturizers Cucumber & Green Tea Aluminum Free with 24-hour Odor Protection 2.6 oz 3 Count
Visit the Dove Store
★★★★☆ 195 ratings | 3 answered questions
Amazon's Choice for "dove 0% aluminum"

Price: $22.47 ($2.88 / Ounce) prime FREE One-Day
& FREE Returns

Pay ~~$22.47~~ $2.13 after using available American Express Membership Rewards points.

Returnable until Jan 31, 2021

Scent	Cucumber,green tea
Brand	Dove
Material type free	Aluminum Free
Item Volume	2.6 Fluid Ounces

About this item

- Dove 0% Aluminum Deodorant Cucumber & Green Tea 2.6 oz. contains 0% aluminum
- The aluminum free deodorant provides 24 hour odor protection
- Women's deodorant that contains ¼ moisturizers for soft and smooth underarms

(b) Aluminum

Dove Advanced Care Antiperspirant For Sweat Block Cool Essentials 48-Hour Deodorant Protection 2.6 oz 4 Count
Visit the Dove Store
★★★★★ 2,966 ratings | 14 answered questions
Amazon's Choice for "dove deodorant"

List Price: ~~$20.99~~
Price: $17.06 ($1.64 / Ounce) prime FREE One-Day
& FREE Returns
You Save: $3.93 (19%)

Pay ~~$17.06~~ $0.00 after using available American Express Membership Rewards points.

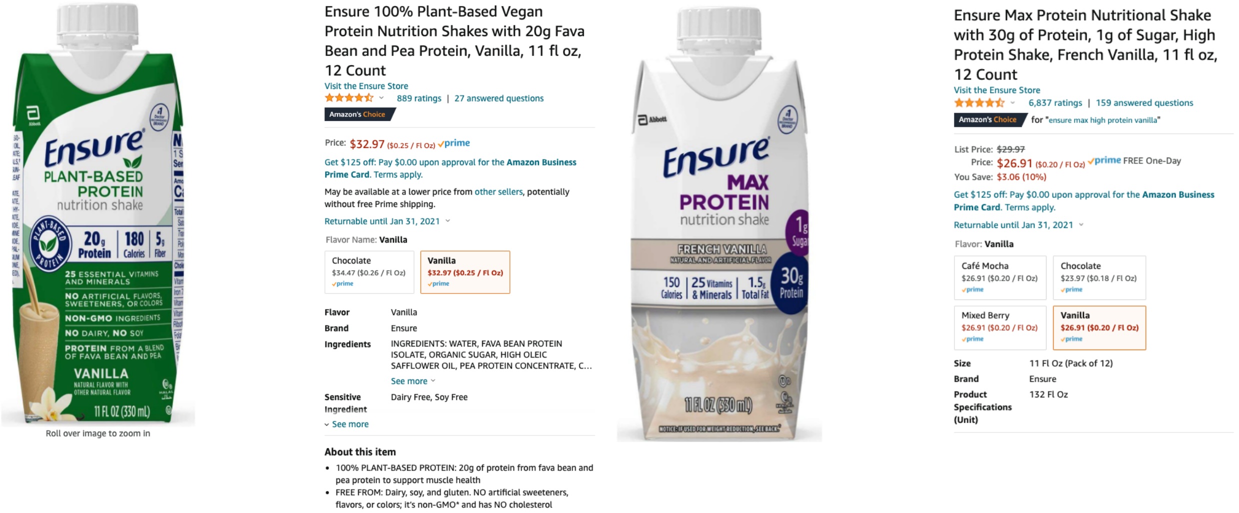
Returnable until Jan 31, 2021

Scent	Green tea, Cucumber
Ingredients	Active Ingredients: Aluminum Zirconium Tetrachlorohydrex Gly (15.2%). Purpose: Anti-Perspirant. Inactive Ingredients: Cyclopentasi… See more
Brand	Dove
Item Form	Stick
Material type free	Alcohol Free

Notes: This figure displays the prices on Amazon for variants of Dove deodorants with and without aluminum. This example was chosen because 1) this brand was included in the aluminum survey, and 2) there exist variants of the same, or similar products with and without aluminum within the same brand. The prices were accessed in November 2020.

[23] AsPredicted #47372, #48205 and #49760, respectively

Figure B3: Product Pricing Examples - Nutritional Shakes with and without GMOs



(a) Non-GMO

Ensure
PLANT-BASED
PROTEIN
nutrition shake
20g Protein
180 Calories
5g Fiber
25 ESSENTIAL VITAMINS AND MINERALS
NO ARTIFICIAL FLAVORS, SWEETENERS, OR COLORS
NON-GMO INGREDIENTS
NO DAIRY, NO SOY
PROTEIN FROM A BLEND OF FAVA BEAN AND PEA
VANILLA
NATURAL FLAVOR WITH OTHER NATURAL FLAVOR
11 FL OZ (330 mL)
Roll over image to zoom in

Ensure 100% Plant-Based Vegan Protein Nutrition Shakes with 20g Fava Bean and Pea Protein, Vanilla, 11 fl oz, 12 Count

Visit the Ensure Store

889 ratings | 27 answered questions

Amazon's Choice

Price: $32.97 ($0.25 / Fl Oz) ✓prime

Get $125 off: Pay $0.00 upon approval for the **Amazon Business Prime Card**. Terms apply.

May be available at a lower price from other sellers, potentially without free Prime shipping.

Returnable until Jan 31, 2021

Flavor Name: **Vanilla**

Chocolate
$34.47 ($0.26 / Fl Oz)
✓prime

Vanilla
$32.97 ($0.25 / Fl Oz)
✓prime

Flavor	Vanilla
Brand	Ensure
Ingredients	INGREDIENTS: WATER, FAVA BEAN PROTEIN ISOLATE, ORGANIC SUGAR, HIGH OLEIC SAFFLOWER OIL, PEA PROTEIN CONCENTRATE, C... See more
Sensitive Ingredient	Dairy Free, Soy Free

See more

About this item

- 100% PLANT-BASED PROTEIN: 20g of protein from fava bean and pea protein to support muscle health
- FREE FROM: Dairy, soy, and gluten. NO artificial sweeteners, flavors, or colors; it's non-GMO* and has NO cholesterol

(b) GMO

Ensure Max Protein Nutritional Shake with 30g of Protein, 1g of Sugar, High Protein Shake, French Vanilla, 11 fl oz, 12 Count

Visit the Ensure Store

6,837 ratings | 159 answered questions

Amazon's Choice for "ensure max high protein vanilla"

List Price: ~~$29.97~~
Price: $26.91 ($0.20 / Fl Oz) ✓prime FREE One-Day
You Save: $3.06 (10%)

Get $125 off: Pay $0.00 upon approval for the **Amazon Business Prime Card**. Terms apply.

Returnable until Jan 31, 2021

Flavor: **Vanilla**

Café Mocha
$26.91 ($0.20 / Fl Oz)
✓prime

Chocolate
$23.97 ($0.18 / Fl Oz)
✓prime

Mixed Berry
$26.91 ($0.20 / Fl Oz)
✓prime

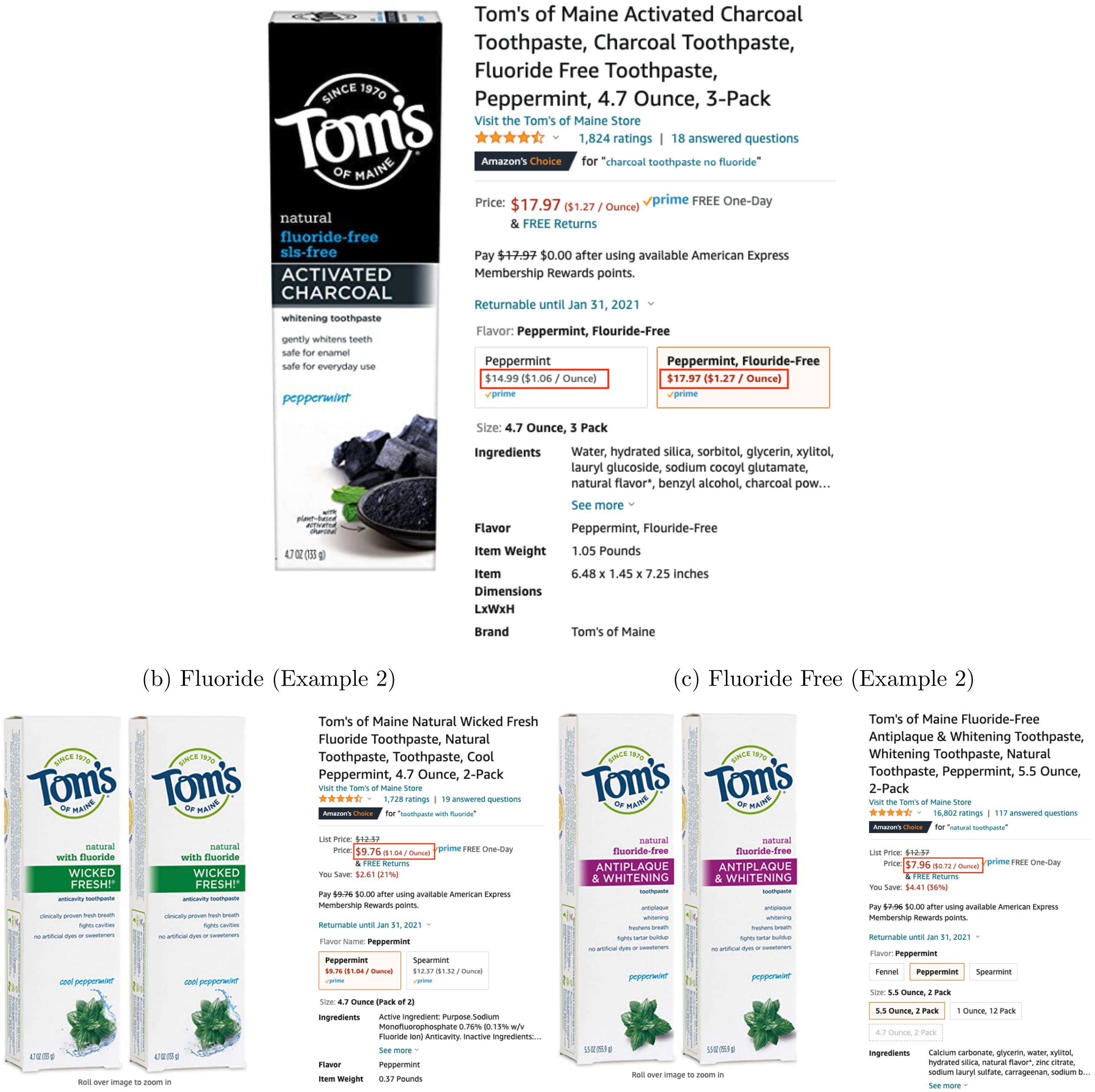
Vanilla
$26.91 ($0.20 / Fl Oz)
✓prime

Size	11 Fl Oz (Pack of 12)
Brand	Ensure
Product Specifications (Unit)	132 Fl Oz

Notes: This figure displays the prices on Amazon for variants of Ensure nutritional shakes with and without GMOs. This example was chosen because 1) this brand was included in the GMO survey, and 2) there exist variants of the same, or similar products with and without GMOs within the same brand. The prices were accessed in November 2020.

Figure B4: Product Pricing Examples - Toothpastes With and Without Fluoride

(a) Flouride and Fluoride-free: Example 1



Tom's of Maine Activated Charcoal Toothpaste, Charcoal Toothpaste, Fluoride Free Toothpaste, Peppermint, 4.7 Ounce, 3-Pack

Visit the Tom's of Maine Store

1,824 ratings | 18 answered questions

Amazon's Choice for "charcoal toothpaste no fluoride"

Price: $17.97 ($1.27 / Ounce) prime FREE One-Day & FREE Returns

Pay ~~$17.97~~ $0.00 after using available American Express Membership Rewards points.

Returnable until Jan 31, 2021

Flavor: **Peppermint, Flouride-Free**

Peppermint $14.99 ($1.06 / Ounce) prime	**Peppermint, Flouride-Free** **$17.97 ($1.27 / Ounce)** prime

Size: **4.7 Ounce, 3 Pack**

Ingredients	Water, hydrated silica, sorbitol, glycerin, xylitol, lauryl glucoside, sodium cocoyl glutamate, natural flavor*, benzyl alcohol, charcoal pow… See more
Flavor	Peppermint, Flouride-Free
Item Weight	1.05 Pounds
Item Dimensions LxWxH	6.48 x 1.45 x 7.25 inches
Brand	Tom's of Maine

(b) Fluoride (Example 2)

Roll over image to zoom in

Tom's of Maine Natural Wicked Fresh Fluoride Toothpaste, Natural Toothpaste, Toothpaste, Cool Peppermint, 4.7 Ounce, 2-Pack

Visit the Tom's of Maine Store

1,728 ratings | 19 answered questions

Amazon's Choice for "toothpaste with fluoride"

List Price: ~~$12.37~~

Price: $9.76 ($1.04 / Ounce) prime FREE One-Day & FREE Returns

You Save: $2.61 (21%)

Pay ~~$9.76~~ $0.00 after using available American Express Membership Rewards points.

Returnable until Jan 31, 2021

Flavor Name: **Peppermint**

Peppermint **$9.76 ($1.04 / Ounce)** prime	Spearmint $12.37 ($1.32 / Ounce) prime

Size: **4.7 Ounce (Pack of 2)**

Ingredients	Active Ingredient: Purpose.Sodium Monofluorophosphate 0.76% (0.13% w/v Fluoride Ion) Anticavity. Inactive Ingredients:… See more
Flavor	Peppermint
Item Weight	0.37 Pounds

(c) Fluoride Free (Example 2)

Roll over image to zoom in

Tom's of Maine Fluoride-Free Antiplaque & Whitening Toothpaste, Whitening Toothpaste, Natural Toothpaste, Peppermint, 5.5 Ounce, 2-Pack

Visit the Tom's of Maine Store

16,802 ratings | 117 answered questions

Amazon's Choice for "natural toothpaste"

List Price: ~~$12.37~~

Price: $7.96 ($0.72 / Ounce) prime FREE One-Day & FREE Returns

You Save: $4.41 (36%)

Pay ~~$7.96~~ $0.00 after using available American Express Membership Rewards points.

Returnable until Jan 31, 2021

Flavor: **Peppermint**

Fennel	**Peppermint**	Spearmint

Size: **5.5 Ounce, 2 Pack**

5.5 Ounce, 2 Pack	1 Ounce, 12 Pack	4.7 Ounce, 2 Pack

Ingredients	Calcium carbonate, glycerin, water, xylitol, hydrated silica, natural flavor*, zinc citrate, sodium lauryl sulfate, carrageenan, sodium b… See more

Notes: This figure displays the prices on Amazon for variants of Tom's of Maine toothpastes with and without fluoride. These examples were chosen because 1) this brand was included in the fluoride survey, and 2) there exist variants of the same, or similar products with and without fluoride within the same brand. Note that there are two examples of flouride vs fluoride-free pricing. For both deodorant and nutritional shakes, the aluminum-free and GMO-free versions are more expensive than the aluminum and GMO variants, respectively. However, for toothpaste, this is not always the case, as highlighted by this example. The prices were accessed in November 2020.

Table B1: Debunking Message Content

Survey	**Debunking Type**	**Message**
Aluminum	Control	Egyptians are often credited with developing the first deodorant, applying sweet-smelling scents to cover up body odor. Their deodorants consisted of spices, such as citrus or cinnamon.
	Treatment	Aluminum-containing products are safe for topical use. Aluminum in deodorant products prevents sweat buildup, and scientific studies have found no conclusive evidence that it causes adverse health effects.
Fluoride	Control	Egyptians are often credited with developing the first toothpaste. The earliest Egyptian recipe contained plenty of abrasives to scrape off all the sticky residue: the ashes of burnt egg shells and oxen hooves mixed with pumice seemed to be popular.
	Treatment	Fluoride-containing toothpastes are safe. Fluoride in toothpastes prevents cavities, and scientific studies have found no conclusive evidence that it causes adverse health effects.
GMO	Control	Whey protein is a nutritional supplement that comes from milk. It's isolated from the rest of the milk through a variety of purification processes. Only 20 percent of milk's protein is whey.
	Treatment	GMOs are safe. GMOs benefit the environment by creating more sustainable farming methods, and scientific studies have found GMO foods are just as safe as non-GMO foods.

Notes: This table displays the debunking messages for all debunking types. The treatment group encompasses the firm, media, and regulator groups, as the debunking messages are the same across all sources. Each debunking message also includes a website link to an actual article from the source.

Table B2: Post-Conjoint Survey Questions: Aluminum

Question Type	Question
Verification Checks	Which account is the first Tweet from? Which account is the second Tweet from? What is the first Tweet's point? What was the second Tweet's point?
Product Usage	In the past week, how frequently have you used deodorant?
Brand Preference	What is your favorite deodorant brand?
Sensitivity to Ingredients	In instances where you look at ingredients when purchasing deodorant, why? Select all that apply.
Knowledge about the ingredient	What is the reason for adding aluminum to deodorants and antiperspirants?
Other controversial products	Would you buy the following? Choose all that apply.
Ingredient Importance	Do you generally read ingredient labels?

Notes: This table includes the questions asked in the aluminum survey, except the conjoint and demographic questions. The demographic questions are the same across all surveys and are displayed in Table B5. All questions are multiple choice.

Table B3: Post-Conjoint Survey Questions: Fluoride

Question Type	Question
Verification Checks	Which account is the first Tweet from? Which account is the second Tweet from? What is the first Tweet's point? Which of the following best describes the second Tweet's point?
Product Usage	In the past week, how frequently have you brushed your teeth?
Brand Preference	What is your favorite toothpaste brand?
Sensitivity to Ingredients	In instances where you look at ingredients when purchasing toothpaste, why? Select all that apply.
Beliefs about the focal ingredient	Do you believe that fluoride in toothpaste is harmful to humans?
Knowledge about the ingredient	What is the reason for adding fluoride to toothpaste?
Other controversial products	Would you buy the following? Choose all that apply.
Ingredient Importance	Do you generally read ingredient labels?

Notes: This table includes the questions asked in the fluoride survey, except the conjoint and demographic questions. The demographic questions are the same across all surveys and are displayed in Table B5. All questions are multiple choice.

Table B4: Post-Conjoint Survey Questions: GMO

Question Type	Question
Verification Checks	Which account is the first Tweet from? Which account is the second Tweet from? What is the first Tweet's point? What was the second Tweet's point?
Product Usage	In the past month, have you purchased any nutritional shakes? Generally, how do you consume nutrition shakes? Choose all that apply.
Brand Preference	Among the options below, what is your favorite brand?
Sensitivity to Ingredients	In instances where you look at ingredients when purchasing a shake, why? Select all that apply.
Beliefs about the focal ingredient	Do you believe that GMOs are harmful to health?
	Do you believe that GMOs benefit the environment?
Knowledge about the ingredient	What is the reason for genetically modifying crops?
Other controversial products	Would you buy the following? Choose all that apply.
Ingredient Importance	Do you generally read ingredient labels?

Notes: This table includes the questions asked in the GMO survey, except the conjoint and demographic questions. The demographic questions are the same across all surveys and are displayed in Table B5. All questions are multiple choice.

Table B5: Demographics Survey Questions

Question Type	Question
Cultural Cognition†	People in our society often disagree about issues of equality and discrimination. How strongly you agree or disagree with each of these statements?
Education	What is your highest level of education achieved?
Household Income	What is your annual household income level?
Race	Choose one or more races that you consider yourself to be:
Political Affiliation	Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else?
Primary Shopper	Are you the primary grocery shopper for your household?
Children	How many children do you have?
Grocery Spend	How much do you spend on grocery shopping per month?

Notes: This table reports the demographics questions that are asked in all of the surveys. All questions are multiple choice.

†: Survey questions are taken from Kahan (2008).

Table B6: Randomization Check - Ad

	Control	Treated	P-value
Aluminum			
N	602	1195	–
Age	33.65	33.23	0.49
Prop. Female	0.53	0.53	0.97
Prop. Democrat	0.5	0.46	0.15
Prop. White	0.75	0.77	0.29
Prop. Black	0.08	0.08	0.89
Fluoride			
N	1619	1583	–
Age	33.12	32.5	0.15
Prop. Female	0.54	0.52	0.2
Prop. Democrat	0.51	0.5	0.92
Prop. White	0.74	0.74	0.99
Prop. Black	0.11	0.11	0.78
GMO			
N	787	772	–
Age	32.45	32.67	0.72
Prop. Female	0.51	0.51	0.99
Prop. Democrat	0.49	0.47	0.48
Prop. White	0.76	0.72	0.06
Prop. Black	0.08	0.09	0.41

Notes: This table reports the demographics for participants in each ad type. The p-value is associated with the null hypothesis that the mean values of the corresponding row is the same across all debunking groups.

Table B7: Randomization Check - Debunking

	Control	Firm	Media	Regulator	P-value
Aluminum					
N	451	438	440	468	–
Age	33.39	33.81	33.15	33.13	0.83
Prop. Female	0.54	0.52	0.54	0.54	0.85
Prop. Democrat	0.43	0.48	0.5	0.49	0.14
Prop. White	0.77	0.76	0.78	0.76	0.88
Prop. Black	0.09	0.07	0.05	0.1	0.1
Fluoride					
N	812	793	784	813	–
Age	32.78	33.47	32.71	32.31	0.3
Prop. Female	0.54	0.52	0.53	0.52	0.64
Prop. Democrat	0.52	0.47	0.56	0.47	0
Prop. White	0.76	0.72	0.74	0.73	0.39
Prop. Black	0.1	0.1	0.12	0.11	0.47
GMO					
N	386	407	382	384	–
Age	32.87	32.14	32.92	32.33	0.75
Prop. Female	0.51	0.52	0.47	0.55	0.14
Prop. Democrat	0.42	0.47	0.51	0.52	0.03
Prop. White	0.76	0.73	0.72	0.75	0.49
Prop. Black	0.09	0.09	0.1	0.08	0.86

Notes: This table reports the demographics for participants in each debunking group. The p-value is associated with the null hypothesis that the mean values of the corresponding row is the same across all debunking groups.

Table B8: Ingredient Surveys Verification Check - Ads

Survey	Ad	Source Correct	Content Correct
Aluminum	Control	0.94	0.93
Aluminum	Treatment	0.92	0.73
Fluoride	Control	0.93	0.96
Fluoride	Treatment	0.97	0.93
GMO	Control	0.97	0.81
GMO	Treatment	0.98	0.94

Notes: This table reports the proportion of the participants that passed the verification check for the ad source and content for each survey. the demographics for participants in each debunking group. The p-value is associated with the null hypothesis that the mean values of the corresponding row is the same across all debunking groups.

Table B9: Ingredient Surveys Verification Check - Debunking Messages

Survey	Debunk Source	Source Correct	Content Correct
Aluminum	Control	0.89	0.93
Aluminum	Firm	0.89	0.93
Aluminum	Media	0.82	0.90
Aluminum	Regulator	0.90	0.93
Fluoride	Control	0.89	0.94
Fluoride	Firm	0.92	0.96
Fluoride	Media	0.89	0.96
Fluoride	Regulator	0.92	0.97
GMO	Control	0.80	0.92
GMO	Firm	0.88	0.94
GMO	Media	0.87	0.93
GMO	Regulator	0.85	0.91

Notes: This table reports the proportion of the participants that passed the verification check for each debunking message source and content for each survey.s the demographics for participants in each debunking group. The p-value is associated with the null hypothesis that the mean values of the corresponding row is the same across all debunking groups.

Table B10: Main Results for Demand Estimates using People Passing Verification Checks

	(1)	(2)	(3)
Ingredient	Aluminum	Fluoride	GMOs
ingredient x controlad	-0.916***	1.584***	-0.668***
	(0.150)	(0.0930)	(0.114)
ingredient x controlad x debunk_firm	0.493**	3.16e-05	0.0422
	(0.211)	(0.126)	(0.145)
ingredient x controlad x debunk_media	0.434**	0.0468	0.103
	(0.214)	(0.127)	(0.156)
ingredient x controlad x debunk_regulator	0.736***	0.0165	0.162
	(0.191)	(0.123)	(0.153)
ingredient x misinfoad	-1.039***	0.878***	-0.803***
	(0.100)	(0.0935)	(0.0998)
ingredient x misinfoad x debunk_firm	0.509***	0.328***	0.215
	(0.153)	(0.121)	(0.137)
ingredient x misinfoad x debunk_media	0.780***	0.547***	0.0211
	(0.146)	(0.129)	(0.144)
ingredient x misinfoad x debunk_regulator	0.995***	0.515***	0.0608
	(0.143)	(0.126)	(0.136)
price	-0.446***	-0.475***	-0.187***
	(0.0188)	(0.0135)	(0.00736)
Control dummies	balancing attributes, brands		
N individual	1,192	2,633	1,125

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered by individuals.

Notes: This table presents robustness results using people in the ingredient surveys who passed the verification checks. Conclusions stay the same except that firm debunking is no longer significant among GMO. We group together the weak and strong versions of misinformed ads in deodorant study and report the original estimation results in robustness check.Balancing attributes are: ”scented” for deodorant, ”whitening” for toothpaste, and ”vanilla/chocolate” for nutrition shakes. Brands included in the studies are: {Kopari, Dove, Speedstick} for deodorant, {Colgate, Crest, Tom’s of Maine, Risewell} for toothpastes, and {Soylent, Orgain, Ensure} for nutrition shakes.

Table B11: Main Results of Demand Estimates for Deodorants

	(1)	(2)
ingredient x controlad	-0.819***	-0.916***
	(0.126)	(0.150)
ingredient x controlad x debunk_firm	0.409**	0.493**
	(0.176)	(0.211)
ingredient x controlad x debunk_media	0.312*	0.435**
	(0.178)	(0.214)
ingredient x controlad x debunk_regulator	0.582***	0.736***
	(0.166)	(0.191)
ingredient x exploitad	-0.954***	-1.119***
	(0.116)	(0.145)
ingredient x exploitad x debunk_firm	0.544***	0.645***
	(0.159)	(0.216)
ingredient x exploitad x debunk_media	0.650***	0.957***
	(0.168)	(0.214)
ingredient x exploitad x debunk_regulator	0.709***	0.963***
	(0.171)	(0.210)
ingredient x spreadad	-0.752***	-0.949***
	(0.112)	(0.138)
ingredient x spreadad x debunk_firm	0.151	0.353
	(0.172)	(0.217)
ingredient x spreadad x debunk_media	0.432***	0.591***
	(0.155)	(0.197)
ingredient x spreadad x debunk_regulator	0.744***	1.010***
	(0.152)	(0.195)
price	-0.426***	-0.446***
	(0.0153)	(0.0188)
Control dummies	balancing attributes, brands	
Everyone Passed all verification checks?	N	Y
N individuals	1,797	1,192

*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors clustered by individuals.

Notes: This table presents the robust result for deodorant when we keep two versions of treated ads separate, in the ingredient surveys. The first column is for all respondents, second column is for only respondents who passed the verification checks. The “exploit” (weak) version of ad states that the deodorant is aluminum-free, whereas the “spread” (strong) version explicitly states that “non-natural” deodorants have toxic ingredients. Balancing attribute for deodorants is: ”Scented”. Brands included in the deodorant study are: Kopari, Dove, Speed Stick.

Table B12: Participant Characteristics in Beliefs Survey

Treatment Group	N	Age	Prop. Female	Prop Democrat	Prop. White	Prop. Black
Treated Ad - Regulator Debunk	1210	35	0.57	0.50	0.74	0.10
Control Ad - Regulator Debunk	1188	34.9	0.58	0.54	0.73	0.09
Treated Ad - Control Debunk	1178	34.3	0.56	0.51	0.75	0.10
Control Ad - Control Debunk	1182	34.8	0.56	0.51	0.73	0.11

Notes: This table reports the demographics for participants in each treatment group in the beliefs survey. The demographics are not statistically different across treatment groups. The p-values from the ANOVA in which the null hypothesis is the mean values of age, proportion female, Democrat, white, black are the same across treatment groups are 0.49, 0.64, 0.3, 0.82, and 0.70, respectively.