Article

AMERICAN MARKETING ASSOCIATION

Journal of Marketing Research
2024, Vol. 61(4) 659-681

DOI: 10.1177/00222437221147088
journals.sagepub.com/home/mrj
Sage

Debunking Misinformation About Consumer Products: Effects on Beliefs and Purchase Behavior

Jessica Fong, Tong Guo, and Anita Rao

Abstract

The prevalence of misinformation has spurred various interested parties—regulators, the media, and competing firms—to debunk false claims in the marketplace. This research examines whether such debunking messages provided by these parties can impact consumer purchase behavior. If so, does debunking effectively correct consumers' misinformed beliefs—an ideal outcome from a policy maker's perspective—or does it merely reinforce correct beliefs, as predicted by biased belief updating? With theory providing contradictory predictions, the authors design and implement a conjoint experiment that enables measurement of 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 deodorant, fluoride in toothpaste, and genetically modified organisms in food), the authors find that debunking plays an important role in mitigating the impact of misinformation. More specifically, debunking can attenuate the decrease in willingness to pay caused by misinformation by correcting misbeliefs, a promising finding for policy makers. The authors discuss the incentives for firms to debunk misinformation or to introduce new products that conform to misinformation.

Keywords

debunking, misinformation, deceptive advertising, social media, beliefs, conjoint, policy

Online supplement: https://doi.org/10.1177/00222437221147088

Submitted December 3, 2021

Misinformation is a widespread issue in the marketplace. Between 2015 and 2020, the Federal Trade Commission (FTC) filed 172 cases regarding misleading advertising and marketing, with settlements up to $191 million (FTC 2019, 2020). Often appearing in the form of unsubstantiated claims that consumers cannot easily verify, such misinformation not only can harm consumers who purchase the product but also can spread misbeliefs about the entire product category, creating negative spillovers into other products. For example, Kopari, a relatively new entrant in the deodorant market, states in a social media post that its product is "aluminum-free" and is therefore nontoxic (Figure 1). Consumers who view this post may form new beliefs that aluminum—a common active ingredient in deodorants and antiperspirants—is harmful, which may increase their willingness to pay (WTP) for aluminum-free products, even though such a claim is not supported by scientific evidence (Palus 2019). The digital era magnifies the urgency of this problem; online material is largely unvetted, making it easy for misinformation to be created and disseminated.

Interested parties have attempted to combat such misinformation through debunking messages. For example, a prominent competitor, Speed Stick, highlights on its website the lack of scientific evidence demonstrating that aluminum in antiperspirants and deodorants is harmful (Duggal 2020). Nevertheless, it has long been debated whether such debunking messages are indeed effective. Moreover, for what type of consumer and through what source(s) can debunking be effective?

This research aims to understand the impacts of misinformation and debunking on purchase behavior. First, can misinformation seen in advertisements influence consumers' WTP by creating misbeliefs? Second, can debunking reduce the impact on WTP caused by misbeliefs created by either advertisements or other sources? Third, does the impact of debunking vary by the source of the message and by the level of consumers' existing (mis)beliefs? We investigate these questions in three

Jessica Fong is Assistant Professor of Marketing, Stephen M. Ross School of Business, University of Michigan, USA (email: jyfong@umich.edu). Tong Guo is Assistant Professor of Marketing, Fuqua School of Business, Duke University, USA (email: tong.guo@duke.edu). Anita Rao is Associate Professor of Marketing, McDonough School of Business, Georgetown University, USA (email: anita.rao@georgetown.edu).

Figure 1. Screenshot of an Instagram Post by Kopari.

product categories in which misinformation about the focal ingredient's harmfulness is prevalent.

From a policy maker's perspective, the ideal outcome is for debunking to correct misinformed beliefs. Whether debunking can achieve such a goal depends on the manner in which individuals update their beliefs. If individuals update in an unbiased Bayesian fashion, then those with prior beliefs most different from the presented information would change their beliefs and their WTP the most. In such cases, debunking can be effective in correcting misinformed beliefs. However, individuals can over- or underweight information as a function of their priors (Schwartzstein 2014) or misinterpret conflicting evidence as favorable to their existing beliefs (Fryer, Harms, and Jackson 2019; Nickerson 1998; Rabin and Schrag 1999). In such cases, debunking would merely confirm existing correct beliefs and not be effective at correcting misinformed beliefs, or even backfire by strengthening misinformed beliefs (Nyhan and Reifler 2010). Therefore, it is important to empirically measure responses along varying levels of existing (mis)beliefs to evaluate the efficacy of debunking.

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., products not containing genetically modified organisms [GMOs]) 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 not only is it impossible to debrief everyone who was exposed to the experimentally induced misinformation (treated consumers might share the experimental ad, making debriefing everyone who was exposed impossible and leading to harmful consequences uncontrollable by the researchers), but it would also be challenging to measure existing beliefs without explicitly asking for them (see Manski 2004). Moreover, if misinformation or debunking does *not* change beliefs in a field experiment, then one cannot identify the reason: are consumers not paying attention to the message, or are they not using the information to update their beliefs even if they do pay attention to the message? Designing a platform to increase individuals' attention to debunking messages (such as providing visual highlights and content recommendations) could be effective in correcting misbeliefs only if the ineffectiveness of debunking is due to inattention.

Therefore, we conduct a controlled online experiment in which it is possible to impose attention checks and to debrief subjects. We measure respondents' true WTP by using 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 not only to recover preferences of the focal ingredient but also to measure how these preferences change under varying exposures to misinformation and debunking.

We define misinformation as any message that does not follow the federal law that states that "an ad must be truthful, not misleading, and, when appropriate, backed by scientific evidence" (FTC 2023). Under this definition, the strength of the misinformation can vary. For example, "ingredient X is toxic" is a stronger statement than "our product is free of ingredient X and therefore is good for you." We consider both to be misinformation if there is no scientific evidence that ingredient X is harmful or if "X-free" is indeed better than "X." We investigate three categories of consumer packaged goods in which misinformation regarding the safety of a main ingredient is prevalent: toothpaste (fluoride), deodorant (aluminum), and nutrition shakes (GMOs), with a separate survey for each ingredient.[1] These product categories also have (1) a firm within the category that circulated messages containing misinformation and (2) a competing firm, regulator, and media that debunk the misinformation. 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 into similar social media posts from arguments presented in actual news articles, on regulatory websites, and on competitors' websites.

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

[1] Basch, Milano, and Hillyer (2019) find that a high proportion of Instagram posts mentioning fluoride contain misinformation. For more information about the misinformation around aluminum in deodorants, see Watson (2011). Regarding GMOs, in 2018, 49% of Americans believed that genetically modified foods are worse for one's health than non–genetically modified foods (Funk, Kennedy, and Hefferon 2018).

conditions: the control group sees an unrelated factoid from "How Stuff Works," and the three treatment groups see debunking messages from the media, a regulator, or a competing firm. By design, the debunking messages across the three sources are identical; they differ only in the source. This design allows us to measure the average impact of misinformation and source-specific debunking on preferences and WTP.

As discussed previously, the efficacy of debunking hinges on the level of responses across prior beliefs, which cannot be measured through choice data alone. For example, a consumer's decision to purchase an aluminum-free deodorant may be driven by preferences (e.g., "I do not like aluminum because it stains clothes") or by beliefs surrounding ingredient toxicity (e.g., "I do not like aluminum because it is toxic"). To measure the extent to which debunking impacts choices by *correcting* misbeliefs, we design an additional experiment at a larger scale that elicits beliefs before and after exposure to the treatments.

Our studies reveal three key findings. First, misinformation in ads generally reduces WTP, and it does so by creating misbeliefs. For consumers with *correct* prior beliefs (i.e., priors inconsistent with misinformation), misinformation significantly decreases WTP: fluoride misinformation decreases WTP by 22% ($.87), and aluminum misinformation decreases WTP for aluminum by 80% ($.48). For consumers with *incorrect* prior beliefs (i.e., priors consistent with misinformation), misinformation has no significant effect. We find no significant effects of GMO misinformation on WTP, likely because this category has the weakest misinformation claim in the study, and because participants are more likely to have been exposed to GMO misinformation, relative to the other categories, prior to the study due to regulator policies. We discuss this in detail in the "Discussion" section.

Second, although debunking can be effective across all sources, debunking by regulators is the only source that has a statistically significant effect for *all* categories: it increases WTP for aluminum, fluoride, and GMOs by 68%, 27%, and 18%, respectively. Debunking by competing firms is effective for fluoride and GMOs, and media debunking is effective for aluminum and fluoride. We explain differences in effects across categories and sources in the "Discussion" section.

Third, we find that debunking can undo the damage caused by both the experimental dose of misinformation and the misbeliefs formed prior to our study. Moreover, for aluminum and GMOs—where most consumers have incorrect prior beliefs—we find that not only does debunking increase WTP, but this increase is significantly *greater* for those with incorrect prior beliefs than for those with correct prior beliefs. We find no evidence that debunking backfires in our context. Rather, we demonstrate that debunking can correct misinformed beliefs, an encouraging finding from a policy maker's perspective.

Finally, motivated by the effectiveness of competitor debunking, we analyze firms' reactions to a new entrant that spreads misinformation. Our estimates allow us to quantify the incentives for incumbents to either debunk or comply with misinformed beliefs by introducing an ingredient-X-free product. Although debunking increases consumers' preferences for the focal ingredient, we show that in equilibrium, each incumbent's best response is to introduce an ingredient-X-free product rather than debunk. This may explain why we commonly see incumbents introducing products that conform to misinformation in the marketplace (e.g., Dove and Speed Stick launched aluminum-free products in 2019 and 2020, respectively).

Related Literature

In this section, we discuss related literature and highlight the contributions of our research. Advertising has been theoretically and empirically well researched. The vast majority of this literature has focused on truthful advertising, and deceptive advertising has only recently received empirical attention. Recent work has focused on review fraud (He, Hollenbeck, and Proserpio 2022; Luca and Zervas 2016; Mayzlin, Dover, and Chevalier 2014) and false claims (Avery et al. 2013; Chiou and Tucker 2018; Kong and Rao 2021; Rao 2022; Rao and Wang 2017).

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 Institutional Review Board requires debriefing anyone exposed to the ad, which is especially challenging 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 (Chiou and Tucker 2018; Rao 2022; Rao and Wang 2017). Although such policy changes provide exogenous variation in the amount of misinformation in the marketplace, such cases are rare. Moreover, we do not know whether the effect of removing misinformation is symmetric to the effect of direct exposure to misinformation. We contribute to this area by measuring the causal effect of the exposure to misinformation on purchase behavior in an incentive-compatible controlled experiment.[2]

This research focuses on investigating the effectiveness of debunking, and the heterogeneity thereof, in combating misinformation.[3] Previous literature suggests that the process of correcting misinformation is complex and remains to be fully understood (Ecker et al. 2022; Lewandowsky et al. 2012, 2015; Schwarz et al. 2007). In advertising studies, most focus on "self-correction" by the firm (as a result of FTC lawsuits) or corrections directly from the FTC, documenting a small but positive effect of corrective advertising on the reduction

[2] 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).

[3] Other efforts in combating misinformation include eliminating the source of misinformation via bans and downvotes (Chiou and Tucker 2018; Pennycook and Rand 2019); nudging, which involves interactions between the message sender and receiver (Lorenz-Spreen et al. 2020; Pennycook, McPhetres, et al. 2020); fact-checking (Guess, Nyhan, and Reifler 2020; Pennycook, Bear, et al. 2020) and, more recently, prebunking (Amazeen, Krishna, and Eschmann 2022).

of stated misbeliefs, both in the lab (Dyer and Kuehl 1974; Mazis and Adkinson 1976) and in the field (Armstrong, Gurol, and Russ 1983; Bernhardt, Kinnear, and Mazis 1986). We contribute to this literature by quantifying the heterogeneous effects of debunking messages on consumer demand across various sources (competitor, media, and regulator) and across consumers' varying levels of existing (mis)beliefs in the marketplace, two dimensions that are directly relevant for policy making. Our demand estimates allow us to simulate the equilibrium outcomes of firms' strategies to combat misinformation, such as debunking or conforming to misinformation.

Existing theories on how consumers update their beliefs provide conflicting predictions regarding whether debunking can effectively correct misbeliefs. On the one hand, consumers may choose to selectively process information, internalizing only the content that agrees with their existing beliefs. Literature on persuasion and biases suggests that corrections that are incompatible with existing (mis)beliefs tend to be processed less fluently (Lewandowsky et al. 2012) and misinterpreted as favorable to existing beliefs (Fryer, Harms, and Jackson 2019; Nickerson 1998; Rabin and Schrag 1999). As a result, information inconsistent with beliefs can be underweighted (Schwartzstein 2014), be ignored, or unintentionally strengthen the consumer's original attitudes (Tormala and Petty 2004). For example, in the context of political misperceptions, Nyhan and Reifler (2010) empirically document the backfire effect, in which debunking messages *increase* misperceptions. Irrespective of the exact mechanism, these theories suggest that debunking is unlikely to correct beliefs for those with prior misbeliefs. On the other hand, an unbiased Bayesian updating framework predicts that individuals update their beliefs when presented with information (from a reputable source) that conflicts with their priors. Recent empirical literature, such as Coutts (2019) and Tappin, Pennycook, and Rand (2020), finds evidence that individuals' posterior beliefs in response to information are consistent with the unbiased Bayesian benchmark. Given mixed evidence across various contexts, it is unclear whether a policy maker with intentions to correct the most misinformed beliefs would be successful in their goals. Therefore, we take this question to data by eliciting beliefs directly before treatment and measuring the WTP for consumers with varying levels of misbeliefs.[4]

Broadly, our work also contributes to the literature on consumer responses to information on nutrition and ingredient labels (see, e.g., Bollinger, Leslie, and Sorensen 2011; Liaukonyte et al. 2013; Scott and Rozin 2020) by studying the impact of misleading information pertaining to various ingredients. This research is also related to a large body of literature in psychology and communications that has analyzed the type of message, the level of detail of the message, and the direction of the cognitive activity of the audience in predicting debunking effectiveness across contexts (see Chan et al. 2017 for a meta-analysis). We refer readers to Walter and Murphy (2018) for a comprehensive review of literature on misinformation and debunking in health, journalism, science, and politics.

In summary, our work extends the literature on debunking misinformation 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 preexisting beliefs, allowing us to comment on the mechanism by which individuals process the corrections to misleading claims. Third, we explore heterogeneity in the effectiveness of debunking by sources commonly observed in practice. To our knowledge, no study has compared the effectiveness of debunking by competitors with that from regulators or mainstream media. Although competitor advertising is more accessible than messages by regulators and media, it may be perceived as a competitive attack on the rival brand, thus carrying little weight in correcting misbeliefs. Other work has shown that competitors' messages can dilute the message of the focal brand (Burke and Srull 1988; Danaher, Bonfrer, and Dhar 2008; Keller 1991). This suggests that, irrespective of the message, the mere presence of the competitive clutter would make the focal ad less effective. In our setting, we measure the specific impact of debunking beyond such clutter, as our control debunking message still adds clutter but does not debunk. Fourth, by experimentally creating variation in exposure to misinformation before debunking, we can evaluate whether debunking can "repair" the change in WTP created through misinformation in ads. Finally, our demand estimates allow us to comment on firms' strategic responses. By simulating incumbents' equilibrium reactions to misinformation, our work provides an explanation for why firms may lack the incentive to debunk.

Consumer Responses to Information: A Framework

In this section, we present a framework of how misinformation and debunking messages about a product's ingredient can influence preferences via beliefs. We use this framework to guide our experimental design, described in the next section. Let consumer i's utility for product j be

$$u_{ij} = \beta_i ing_j + \gamma Z_j + \varepsilon_{ij}, \tag{1}$$

where $ing_j \in \{0, 1\}$ is an indicator for whether j contains the focal ingredient, Z_j is a vector consisting of the product's other attributes (such as brand, price, packaging, flavor, scent, etc.), and ε_{ij} is an

[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 modeled the typical assumptions used in belief updating theories with applications to heterogeneous price search and brand choices (Charness and Levin 2005; Ching et al. 2021; Jindal and Aribarg 2021; Nyarko and Schotter 2002; Ursu et al. 2023).

idiosyncratic shock that varies at the consumer-product level. The coefficient of interest is β_i, that is, i's preference for the focal ingredient. For example, if the product j is deodorant, then $ing_j = 1$ indicates that the deodorant contains aluminum, and β_i is i's preference for aluminum in deodorant.

Information 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. The closer θ_i is to 1, the more certain i is that the ingredient is harmful. Similarly, the closer θ_i is to 0, the more certain i is that the ingredient is *not* harmful. The coefficient δ represents the population-average risk attitude toward the harmfulness of the ingredient; δ is likely negative, as people are typically averse to ingredients they believe to be harmful. The variable τ is the population-average inherent preference for the ingredient, which can be either positive or negative.[5] Finally, η_i are individual-level differences in the preferences for the ingredient that are unexplained by θ_i.

Debunking messages and misinformation about an ingredient can impact individual i's utility for j by changing θ_i, their belief about whether the ingredient is harmful. The sequence of events is straightforward: individual i has a prior belief θ_i^0, receives information about the ingredient ("ingredient X is harmful" or "ingredient X is not harmful") from a source, such as a regulator, and then updates their beliefs to θ_i^{post}. However, the extent to which individuals update their beliefs in response to new information is not as straightforward. Consumers can biasedly choose whether to use new information to update beliefs according to their prior beliefs and/or the credibility of the source of the information. Subsequently, we discuss responses to debunking messages and note that the same rationale applies to responses to misinformation.

On one extreme, consumers may choose to selectively process information, internalizing only the content that aligns with their existing beliefs. The literature on confirmation bias points out that individuals can seek out information that is consistent with their hypothesis, overweight evidence that confirms their hypothesis, and underweight evidence that conflicts with their hypothesis (Nickerson 1998). Similarly, the literature on resistance to persuasion documents various reasons why contradictory information can either have no effect or unintentionally strengthen the consumer's original attitudes (Tormala and Petty 2004). In economics, Rabin and Schrag (1999) model confirmation bias as the agent misreading information that conflicts with the agent's priors as information that *supports* the priors. Irrespective of the exact mechanism, these theories suggest that a debunking message is the least likely to correct the beliefs of those with incorrect priors. We demonstrate this using a simple model in Web Appendix A. In this model, biased updating manifests in the form of consumers viewing the source to be less trustworthy if it conveys information that conflicts with the consumer's priors. Such a model predicts that, on average, consumers with incorrect priors update *less* in response to debunking than those with correct priors.

On the other end, an unbiased Bayesian framework posits that a consumer updates their beliefs when presented with new information, and contradicting information (from a reputable source) shifts beliefs the most. For example, Tappin, Pennycook, and Rand (2020) find that individuals' posterior beliefs about political topics in response to new information are close to the unbiased Bayesian benchmark. Under this framework, those with the most misinformed beliefs have the most to learn from the debunking message if the message is perceived as truthful and credible. In Web Appendix A, we also present a model of unbiased Bayesian updating. This model predicts that, conditional on the source being trustworthy, consumers with incorrect priors update *more* in response to debunking than those with correct priors.

To summarize, the two frameworks discussed previously make the following predictions for individuals' responses to debunking.

Prediction 1 (unbiased Bayesian updating): Individuals who believe the ingredient to be harmful (i.e., with priors $\theta^0 > .5$) update their beliefs the most in response to debunking from a trustworthy source. Debunking reduces misbeliefs.[6]

Prediction 2 (biased updating): Individuals who believe the ingredient is not harmful (i.e., with priors $\theta^0 < .5$) update their beliefs the most in response to debunking from a trustworthy source. Debunking only works by further strengthening beliefs for those with already correct beliefs.[7]

Additionally, debunking and misinformation can also be ineffective if the information comes from an untrustworthy source or if the information is consistent with their prior beliefs. We summarize the potential mechanisms for different effects of misinformation and debunking in Tables 1 and 2. Given diverging theoretical predictions and mixed empirical evidence across contexts in existing literature, the empirical effect of misinformation and debunking on beliefs and purchase behavior in this setting is ex ante unclear.

[5] 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 cavities. They may prefer nutrition shakes made from GMOs because GMOs have a smaller environmental impact.

[6] Similarly, when exposed to misinformation, those who believe the ingredient to be relatively harmless (i.e., with priors) will be the most impacted.

[7] When exposed to misinformation, individuals with priors $\theta^0 > .5$ update their priors the most in response to misinformation.

Table 1. Misinformation Effects: Summary and Mechanisms.

Average Effect on Preference	Potential Mechanisms
Negative	Creates new misinformed beliefs
	Strengthens existing misbeliefs
Null	Message is consistent with prior beliefs (ceiling effect)
	Misinformation comes from an untrustworthy source
Positive	Backfire effect; confirmation bias

Table 2. Debunking Effects: Summary and Mechanisms.

Average Effect on Preference	Potential Mechanisms
Negative	Backfire effect; confirmation bias
Null	Message is consistent with prior beliefs (ceiling effect)
	Debunking comes from an untrustworthy source
Positive	Corrects misinformed beliefs
	Strengthens existing correct beliefs

Experiment Design

The first question of this research is the following: Does misinformation and debunking, on average, impact consumers' preferences? To answer this question, we implement an incentive-compatible choice-based conjoint experiment that enforces consumers' attention to the information treatments, separately measuring the effect of misinformation in advertising and the effect of debunking on consumer preferences. In the first set of experiments, which we refer to as the ingredient studies, described in the following sections, 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. We also implement a second set of experiments, referred to as the belief studies, which utilize a similar experimental design as the first set of studies, to explore the heterogeneity in responses to debunking across different prior beliefs. We describe the belief studies in the "Misinformation and Debunking Treatment Effects" section.

Three decisions were necessary, related to (1) the choice of product categories, (2) implementation of treatment conditions, and (3) the method of capturing the outcome of interest (i.e., purchase and WTP). We outline the details informing each of these decisions in the following sections.

Choice of Product Categories

We identified categories in which firms market certain products as "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 in which 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, and (3) nutrition shakes and GMOs. These ingredients remain controversial in the United States, despite no scientific evidence of harm (see Centers for Disease Control and Prevention [CDC] 2020; Penn Medicine 2019). In the deodorant category, Kopari states that its deodorants are aluminum-free and implies that other deodorants containing aluminum are toxic (Figure 2, Column 1). 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 that its toothpastes are fluoride-free and encourages consumers to avoid the toxic ingredients found in traditional fluoride toothpastes (Figure 2, Column 2), whereas 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. In the nutrition shake category, Orgain highlights that its products are GMO-free and thus contain only the "good stuff," implying that products with GMOs are "bad" (Figure 2, Column 3). Competitors (Soylent), the media (NBC), and the regulator (Food and Drug Administration [FDA]) have pointed out that genetically modified plants not only are safe to consume but also can benefit the environment.

Treatment

The ingredient studies adopt a 2 (control ad, treatment ad) by 4 (control, regulator debunking, media debunking, competitor debunking) design, in which all ads and debunking messages are displayed as tweets because the original ads containing misinformation were displayed on social media.[8] Participants are first randomized into receiving either a control ad or a treatment ad. Figure 2 displays the control (top row) and treatment (bottom row) ads across all three products. All ads are real content from the company (see Table W7 in Web Appendix F for the message text). Although the content is not tagged as ads on the platform, we refer to such content as "ads" because they represent messaging that highlights the firm's products, is aimed at consumers, and comes directly from the brand. Participants randomized into the treatment ad group receive an ad that contains misinformation about the focal ingredient,

[8] 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 across categories, 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 that other aspects of the ad, such as the picture, time stamp, and likes and retweets for both ads, are identical across control and treatment conditions.

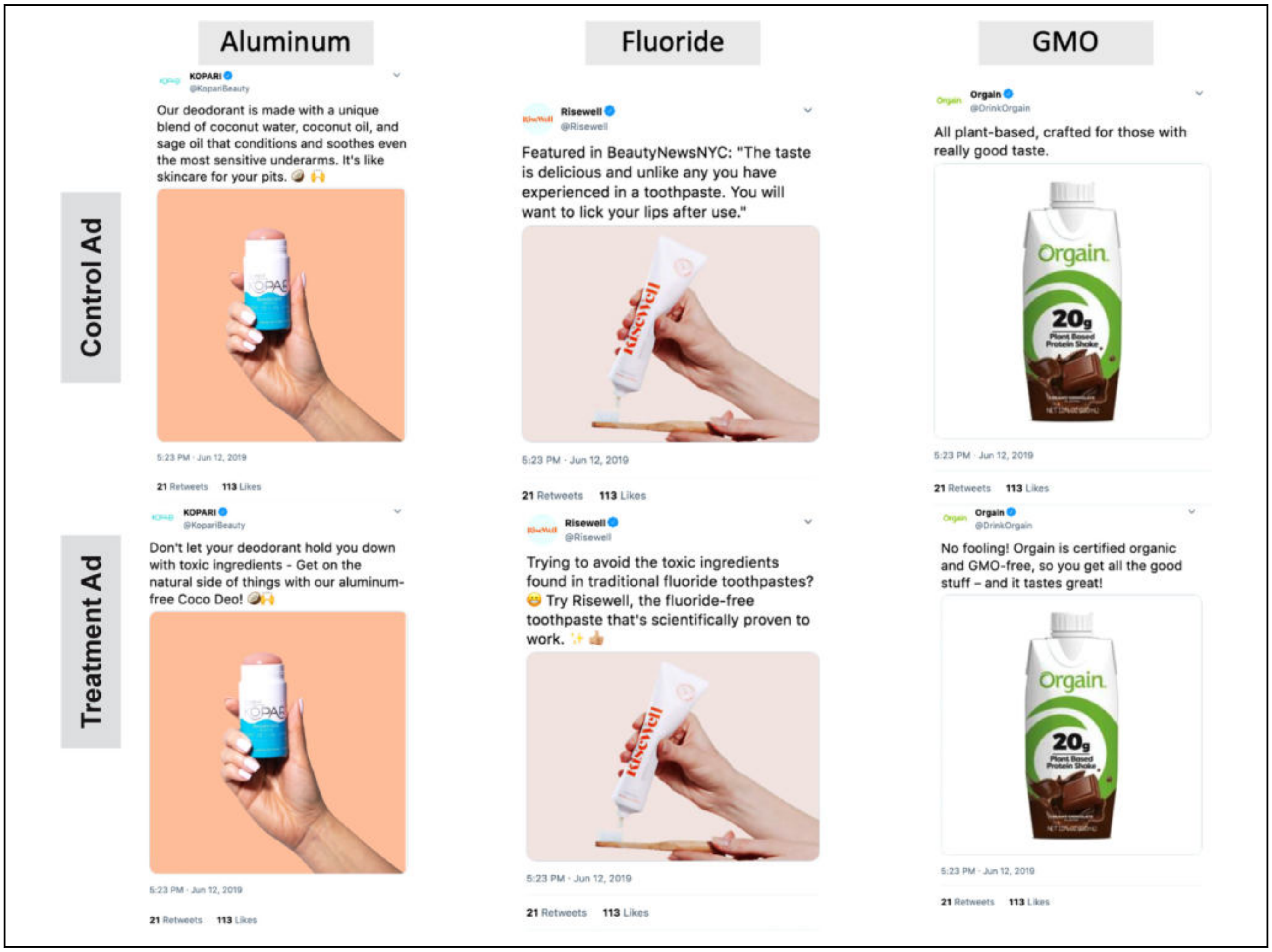

Figure 2. Ads in the Ingredient Studies.
Notes: See Table W7 in Web Appendix F for ad message text. All ad messages are real.

whereas those randomized to the control group receive an ad for the same product that does not mention the focal ingredient. Because we prioritized using actual content from the firms, the control ads are not identical to the treatment ads in all other aspects except for the presence of misinformation. The selected control ads highlight another product attribute (e.g., scent, taste, plant-based formulation). For such ads to be valid controls, we assume that preference for the attribute in the control ad is orthogonal to the preference for the focal ingredient (our main outcome of interest). In other words, we assume that preferences for the highlighted product attribute (e.g., scent, flavor) are not correlated with the preference for the focal ingredient.[9] We note that this assumption is likely valid for fluoride and GMOs: the presence of fluoride does not influence taste, and the presence of GMOs does not influence whether the product is plant-based. However, for aluminum, the control ad highlights the "soothing quality" of the ingredients; if this causes consumers to place more weight on the product's "soothing quality" *and* consumers believe that aluminum can cause skin irritation, then this may decrease preferences for aluminum in the control group. In this case, our measured misinformation treatment effect would be the lower bound of the true misinformation effect. As a robustness check for aluminum, we use a control ad that mentions that the product is aluminum-free but does not contain any misinformation.[10]

[9] We do not include the highlighted attribute in the conjoint because it was not practical to do so. For instance, Risewell emphasizes in the control ad that its toothpastes taste delicious, which is too subjective and intangible as a conjoint attribute. In the absence of the highlighted attribute in the conjoint, preferences for the highlighted attribute are absorbed into preference for the brand through the brand fixed effects and are orthogonal to the focal ingredient. In a robustness check, we control for brand–treatment interactions. The results are displayed in Web Appendix F Tables W20–W21.

[10] This ad is displayed in Web Appendix F Figure W9. The results, reported in Table W19 in Web Appendix F, are robust, thus validating our main control ad for aluminum.

Table 3. Debunking Message Content.

Focal Ingredient	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.

Figure 3. Example of a Debunking Message (Fluoride).

Participants are then randomized into one of the following debunking sources: control, competitor, 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. These debunking messages are summarized from actual articles across all sources. The debunking messages for all categories are presented in Table 3.

Figure 3 presents an example of the debunking message for fluoride, as seen by the participant. The message is accompanied by a link that leads to a real article from the source containing the debunking message. To ensure that only the source varies across all treatment arms, we hold the content of the debunking message constant. The control debunking message is a factoid about the product category that contains no information about the focal ingredient. This factoid is presented as a tweet from the website How Stuff Works. In the remainder of this article, we refer to the control debunking group as the "No Debunking" group. After the ad and treatment exposures, we conduct verification checks on whether the participant can recall the source of the ad and of the debunking message.

Table 4. Categories, Firms Making Misleading Claims, and Debunking Sources.

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

Table 4 reports the category and firm making the false claims, and the debunking sources used in this study. Note that not all sources are held constant within the category (e.g., CDC and FDA). This is because we were unable to find an actual article that debunks the misinformation about the focal ingredient by the same source across ingredients. For instance, we were unable to find an article from the CDC that debunks misinformation about GMOs; therefore, we used an article from the FDA instead.

Our experimental 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 Debunking" and "Misinformation Ad + No Debunking" groups because participants in these two groups differ only by the ad content they were exposed to. Debunking effects after exposure to misinformation for a given source can be measured by comparing those in the "Misinformation Ad + Debunking" group for the given source with those in the "Misinformation Ad + No Debunking" group. The 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 + Debunking" and "Control Ad + No Debunking" groups. Lastly, comparing

Table 5. Product Attributes in Conjoint.

Product	Brand	Has Ingredient	Price ($)	Other Attribute
Deodorant	Dove, Speed Stick, Kopari[a]	Has aluminum: yes/no	1.99, 2.99, 3.99	Scented: yes/no
Toothpaste	Colgate, Crest, Risewell,[a] Tom's of Maine	Has fluoride: yes/no	.99, 1.99, 2.99	Whitens teeth: yes/no
Nutrition shake	Ensure, Orgain,[a] Soylent	GMO-free: yes/no	1.00, 1.25, 1.50[b]	Flavor: chocolate/vanilla

[a]Denotes the advertised brand.
[b]Price per bottle. Due to the logistics of reward distributions, lottery winners in the GMO study received a dozen nutrition shakes, so the conjoint selections are for a pack of a dozen shakes.

participants in the "Control Ad + No Debunking" group with participants in the "Misinformation Ad + Debunking" group reveals the effect of exposure to misinformation *and* debunking (net effect).

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, Grewal, and Liechty (2005) for a discussion on incentive-aligned conjoint analysis.

After exposure to the two treatment conditions, participants are presented with ten conjoint choice tasks. Participants are asked to choose a product from three 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 nutrition shakes). Figure W13 in Web Appendix F provides an example of the choice task faced by participants in the fluoride study. Table 5 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.[11] If they win the lottery, they receive the product that they selected for the given price and the remaining $10 minus the selected price as additional payment. For example, if a participant wins the bonus and had selected a Crest toothpaste with whitening and fluoride for $.99, the participant receives a Crest toothpaste with whitening and fluoride for $.99 and the remaining $9.01 as an additional cash payment. Web Appendix B includes more details on the conjoint instructions for participants and selection of product configurations.

After the conjoint choices, we again conduct verification checks by asking participants to recall the *content* of both the ad and the debunking message. These verification checks are placed after the conjoint questions instead of immediately after the treatments to avoid "treating" the respondents by the choices presented in these verification checks. We also collect information about participants' usage of the product, their opinions about the focal ingredient, and demographics. Tables W9–W12 in Web Appendix F list all the questions asked in this section of the survey. After survey completion, every participant is debriefed with verified scientific content and is informed that the content of the messages may have been modified from the original source.

Misinformation and Debunking Treatment Effects

In this section, we first detail the implementation of the experiments, and present the results from the ingredient studies on the average effect of misinformation and debunking, followed by the debunking effect across sources. We then describe the belief studies, which allow us to measure the heterogeneous effects by prior beliefs. We summarize the main qualitative findings in tables presented in the "Discussion" section.

Ingredient Studies: Average Effects by Ingredient and Source

Implementation and data description. All studies were distributed through Prolific, an online platform for survey administration and data collection. The ingredient studies were launched sequentially in September and October 2020.[12] Such a sequential launch enabled us to exclude participants who had already taken any of our previous surveys. We explicitly did so to avoid any possibility of familiarity with the study and treatment conditions.

The participant pool for each ingredient study was limited to those in the United States and those who did not participate in any of the other studies in this research. Participants received U.S. $1.50 for completing the study, which takes approximately 10 minutes. Our preference elicitation is incentive compatible (see the "Experiment Design" section for implementation details).

In total, 6,558 individuals completed the ingredient studies, with 1,193 participants in the aluminum study, 3,202 in the fluoride study, and 1,559 in the GMO study.[13] The sample

[11] For the GMO study, this was a 1-in-20 chance to win a bonus worth $20 to accommodate the higher price of the dozen pack of nutrition shakes.

[12] The studies were launched on three Wednesdays: September 9, September 23, and October 21, 2020.

[13] For the aluminum study, we tested two versions of control ads: a control ad that highlights another attribute and a control ad that mentions that the product is aluminum-free but does not contain misinformation. We use the control group that highlights another attribute in the main analysis (similar to the other products) and report treatment effects using the latter ad as the control in Web Appendix F. Including the second control ad, the total number of participants in the aluminum study is 1,797.

Table 6. Sample Size for Each Treatment Group.

Ad	Debunk	Aluminum	Fluoride	GMO
Control	Control	155	401	171
Control	Competitor	146	407	217
Control	Media	140	405	195
Control	Regulator	161	406	204
Treatment	Control	140	411	215
Treatment	Competitor	130	386	190
Treatment	Media	155	379	187
Treatment	Regulator	166	407	180

Table 7. Choice Share for Products with the Focal Ingredient.

	Control Ad, No Debunking	Misinformation Ad, No Debunking	Misinformation Ad, Debunking
Aluminum	.29	.28	.39
Fluoride	.67	.56	.64
GMO	.31	.29	.31

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

sizes, which differ across the ingredients, were determined on the basis of the pilot data for each study, as described in Web Appendix E. The three studies are preregistered (https://aspredicted.org/de6c9.pdf, https://aspredicted.org/t53yt.pdf, and https://aspredicted.org/4wg4a.pdf, respectively). Table 6 displays the sample size for each treatment group for each study.

Although the participant pool is similar to the U.S. population in terms of gender and race, the study participants tend to be younger, more educated, and less likely to be unemployed compared with the general U.S. population, as shown in Table W8 in Web Appendix F. Additionally, a higher proportion of study participants self-identify as Democrats relative to the U.S. population. Randomization checks for covariate balance across treatment groups are reported in Tables W13 and W14 in Web Appendix F.

The title of each survey stated the product category explicitly, the goal being to recruit participants interested in and familiar with that category. The findings show that 67% of participants report using deodorant daily, 96% of participants brush their teeth at least once daily, and 36% of participants in the GMO study report purchasing nutrition shakes in the month prior to the respective studies. Additionally, the vast majority of participants passed the verification checks for both the ad and debunking sources and content: 95% (89%) of participants answered the ad (debunking) source verification check correctly, and 89% (94%) answered the ad (debunking) content verification check correctly. We do not see systematic patterns in which sources have lower or higher pass rates across the surveys. Tables W15 and W16 in Web Appendix F report the proportion of participants passing the verification checks by treatment group. As a robustness check, we estimate treatment effects on WTP for both the entire sample and only those who passed the verification checks.

Table 7 displays the share of selected options in the conjoint questions for which the product contains the focal ingredient across various treatment conditions. First, in the control condition (Table 7, Column 1), 29% of the selected options contained aluminum, 67% contained fluoride, and 31% contained GMOs. This suggests that preexisting preferences for or against these ingredients vary across products. In general, consumers avoid aluminum in deodorant and GMOs in food but prefer fluoride in toothpaste.

Second, we find suggestive evidence of the effect of misinformation; the share of options with the focal ingredient is lower among individuals who are exposed to treatment ads containing misinformation (Table 7, Column 2). Among the group exposed to debunking after misinformation (Table 7, Column 3), such a share of options with the focal ingredient is the same as, if not higher than, the control group. In the following section, we formally estimate the treatment effects, controlling for brand and price effects.

Estimation. As described in our theoretical framework, misinformation and debunking can each have a positive, null, or negative impact on preferences. In the results from the ingredient studies, we estimate and document the overall effect of both misinformation and debunking in our setting. Recall that the decision process in the experiment is as follows. After exposure to an ad (control or treatment) and a debunking message from a randomly chosen source (control, competitor, media, or regulator), each individual i is presented with ten sets of product profiles. In each set J, the individual compares three product profiles and the "none" option, and then chooses the one that gives them the highest utility in that set. Formally, the probability of individual i choosing product profile j from set J is

$$\Pr(j)_i = \frac{e^{v_{ij}}}{\sum_{k \in J} e^{v_{ik}}}, \tag{3}$$

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, ing_j, and dummies of exposure to control versus treatment ad (I_i^C, I_i^T) and dummies of debunking messages from a given source s (I_i^s, s ∈ {control, competitor, media, regulator}):

$$\begin{aligned} u_{ij} &= v_{ij} + \varepsilon_{ij} \\ &= \sum_s \beta_1^s ing_j I_i^C I_i^s + \sum_s \beta_2^s ing_j I_i^T I_i^s + \alpha price_j + \gamma Z_j + \varepsilon_{ij}. \end{aligned} \tag{4}$$

We control for preferences for brands and other balancing attributes via brand fixed effects and balancing attribute fixed effects in Z_j. The term ε is assumed to be i.i.d. and has an extreme value type I distribution. The variable β_1^s captures the average preference for the debated ingredient (i.e., aluminum for deodorant, fluoride for toothpaste, and GMOs for nutrition shakes) under debunking source s across participants in the

Table 8. Ingredient Studies: Estimates of Equation 4.

	(1) Deodorant	(2) Toothpaste	(3) Nutrition Shake
Ingredient × control ad	−.815***	1.426***	−.614***
	(.126)	(.0792)	(.0893)
Ingredient × control ad × competitor debunking	.407**	−.00417	.0197
	(.176)	(.109)	(.118)
Ingredient × control ad × media debunking	.311*	.103	.0569
	(.178)	(.110)	(.121)
Ingredient × control ad × regulator debunking	.580***	.0327	.0554
	(.166)	(.109)	(.122)
Ingredient × misinformation ad	−.749***	.829***	−.723***
	(.112)	(.0806)	(.0823)
Ingredient × misinformation ad × competitor debunking	.150	.349***	.230**
	(.171)	(.108)	(.116)
Ingredient × misinformation ad × media debunking	.431***	.521***	−.0223
	(.155)	(.112)	(.121)
Ingredient × misinformation ad × regulator debunking	.741***	.506***	−.0148
	(.151)	(.112)	(.117)
Price	−.414***	−.436***	−.171***
	(.0187)	(.0121)	(.00622)
Balancing attribute	.440***	1.244***	.316***
	(.038)	(.026)	(.038)
Brand dummy			
Brand 1	2.292***	.364***	3.161***
	(.084)	(.050)	(.104)
Brand 2	2.197***	.402***	3.389***
	(.084)	(.049)	(.104)
Brand 3	1.813***	.126***	3.055***
	(.087)	(.053)	(.105)
Brand 4		.197***	
		(.052)	
Ingredient	Aluminum	Fluoride	GMOs
N	1,193	3,202	1,559

*$p < .1$, **$p < .05$, ***$p < .01$.

Notes: This table presents the full model estimates of Equation 4. Balancing attributes are "scented" for deodorant, "whitening" for toothpaste, and "chocolate" for nutrition shakes. For the deodorant study (Column 1), Brands 1, 2, and 3 are Dove, Kopari, and Speed Stick, respectively. For toothpaste (Column 2), Brands 1, 2, 3, and 4 are Colgate, Crest, Risewell, and Tom's of Maine. For nutrition shakes (Column 3), Brands 1, 2, and 3 are Ensure, Orgain, and Soylent. Table W17 in Web Appendix F displays the results for only those who passed the verification checks. Robust standard errors clustered by individuals are in parentheses.

control ad condition, whereas 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.

Results of ingredient studies. Table 8 displays the estimates of Equation 4 for all product categories. In the subsequent sections, we discuss these results, focusing on the effect of the treatments on the WTP of the focal ingredient. Focusing on the WTP as opposed to the ingredient coefficient enables comparison of effect sizes across products. We calculate the WTP using the estimates from the choice model by dividing the ingredient coefficient by the absolute value of the price coefficient.

Baseline preferences. Our empirical results show that baseline preferences (Table 8, Row 1) vary across the three product categories: respondents exhibit a positive WTP for fluoride in toothpastes and a negative WTP for GMOs in shakes and aluminum in deodorants. The average WTPs for aluminum, fluoride and GMOs are −\$1.97 (Table 8, Column 1; −.815/.414), \$3.27 (Table 8, Column 2; 1.426/.436) and −\$3.60 (Table 8, Column 3; −.614/.171), respectively. In other words, prior to the experimental manipulation, participants are on average averse to aluminum and GMOs and have a strong preference for fluoride in toothpastes. Figure 4, Panel A, plots the baseline WTP for all three ingredients.[14]

[14] For comparison, the real-world price differences in November 2020 were the following: a 2.6 oz. Dove aluminum-free deodorant was priced \$2.60 higher than the version with aluminum, a 12-pack of Ensure protein shakes without GMOs was priced \$6.06 higher than the version with GMOs, and a 4 oz. Tom's of Maine toothpaste with fluoride was priced \$1.28 higher than the version without fluoride. As of November 2020 on Amazon, a 2.6 oz. aluminum-free Dove deodorant was \$7.49, and a 2.6 oz. deodorant with aluminum was \$4.89; the Ensure shake without GMO was \$.25 per fl. oz., and the Ensure shake with GMOs was \$.20/fl. oz; the Tom's of Maine peppermint toothpaste without fluoride was \$.72/oz., and the one with fluoride was \$1.04/oz.

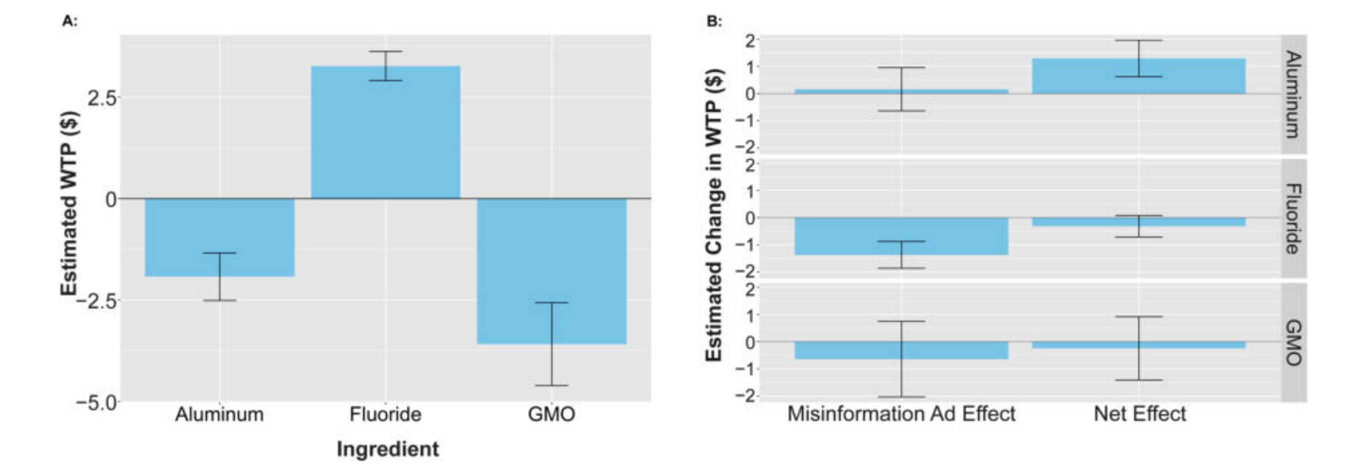

Figure 4. Ingredient Studies: Baseline and Treatment Effects.
Notes: Panel A reports the baseline WTP estimates for aluminum (2.7 oz. deodorants), fluoride (4 oz. toothpastes), and GMOs (a pack of 12 nutrition shakes). Estimates are obtained from the "Control Ad + No Debunking" condition, with price, brands, and other attributes controlled for. Panel B reports the average change in the WTP for different treatment conditions. "Misinformation Ad Effect" is obtained by comparing the "Misinformation Ad + No Debunking" condition with the "Control Ad + No Debunking" condition. "Net Effect" is obtained by comparing the "Misinformation Ad + Competitor/Media/Regulator Debunking" condition with the "Control Ad + No Debunking" condition. Error bars represent the 95% confidence intervals.

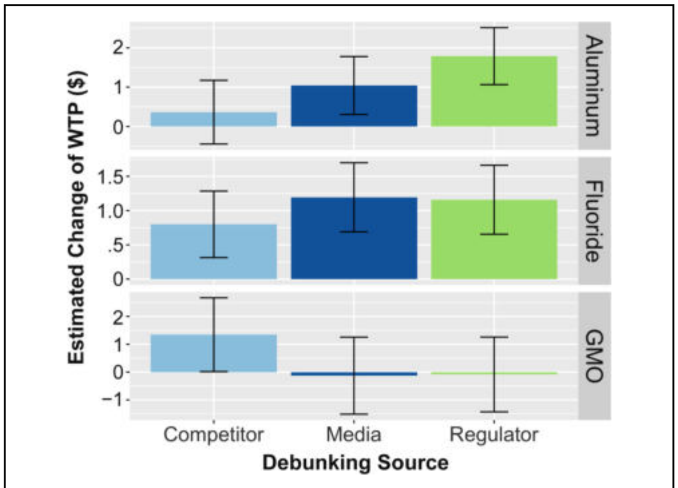

Figure 5. Estimated Debunking Effects on WTP by Source and Ingredient.
Notes: This figure displays the average change in the WTP from different debunking sources for participants who saw an ad with misinformation. The comparison baseline is "Misinformation Ad + No Debunking." Error bars represent the 95% confidence intervals. For GMOs, we found no effect from media or regulator debunking due to a lack of statistical power in detecting the lower-than-expected GMO effect in the ingredient studies. We further increase the sample size and test this in the belief studies.

Effect of misinformation. We find that misinformation can reduce consumers' WTP for the focal ingredient: Figure 4, Panel B (Column 1), shows that, for fluoride, the ad containing misinformation causes a statistically significant $1.37 (42%) decrease in WTP, but the effect is not statistically significant for aluminum or GMOs. Note that prior to being treated with misinformation, consumers are negatively inclined toward aluminum and GMOs but have a strong preference for fluoride (Figure 4, Panel A). As we confirm in the "Results of Belief Studies" section, an absence of response to misinformation does not imply that consumers are immune to misinformation: 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. We find that debunking misinformation increases WTP for the focal ingredient: the first column of Figure 4, Panel B, 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 (first column in Figure 4, Panel B) reverts to zero after debunking (second column in Figure 4, Panel B). The net effect for aluminum is significantly positive despite there being no effect of misinformation, indicating that debunking can impact preexisting preferences. For GMOs, the debunking effect, averaged over all sources, is not statistically significant. Overall, debunking can increase consumers' WTP for aluminum by 65% by impacting preexisting preferences and revert the negative impact of an additional exposure to misinformation for fluoride.

Debunking effectiveness by source. Next, we compare the effectiveness of debunking across the sources: competing firm, media, and regulator, displayed in Figure 5. In each of the product categories, we find at least one source to be effective at debunking misbeliefs, but the differences across sources are not statistically significant. We note that the lack of significance in the differences across the sources could be due to an insufficient sample size to precisely detect differences between the effects. Nevertheless, all three sources of debunking are significantly effective at increasing WTP for fluoride, and two of the three sources are effective at increasing WTP for aluminum. For GMOs, the debunking effects by source are noisily measured; however, the debunking effect is statistically significant for competitors. In the next set of experiments (belief study), we investigate these effects with a larger sample size.

Belief Studies: Heterogeneity by Prior Beliefs

With the first set of experiments measuring the average effect of misinformation and debunking across multiple sources, we next turn to investigating the heterogeneity in effectiveness of debunking by prior beliefs. The primary goal is to understand whether debunking impacts those with the most misinformed beliefs (an ideal outcome for a social planner trying to reduce harm) or whether it merely impacts those who already believe the ingredient not to be harmful. To do so, we implement a second set of studies (belief studies) in which we explicitly elicit participants' prior and posterior beliefs, in addition to their revealed preferences.

Eliciting beliefs. The direct elicitation of beliefs is important because preferences cannot be separately identified from beliefs using choice data alone without imposing more structure on the belief updating process, as illustrated by Equation 2.[15] Both preferences (τ) and beliefs (θ) can lead to consumers not choosing a product with the focal ingredient. For example, a consumer who believes aluminum is not harmful ($\theta = 0$) may still dislike aluminum because it stains clothes ($\tau < 0$) and thus choose to buy an aluminum-free deodorant. Another consumer may believe that aluminum is toxic ($\theta = 1$) and thus choose to buy an aluminum-free deodorant. Manski (2004) suggests directly eliciting beliefs from respondents to circumvent this issue.

The design of the belief studies is identical to that of the ingredient studies except for three modifications. First, we elicit consumers' beliefs about ingredient toxicity before and after responding to the choice questions.[16] To avoid creating a demand effect (in which consumers see the belief questions about the ingredient and infer that this study has something to do with ingredient toxicity and change their responses to satisfy the researcher), we also elicit their beliefs for all other attributes (e.g., brand and whitening) included in the study. We elicit beliefs using the probabilistic elicitation method suggested by Manski (2004). Before and after the choice questions, participants are asked: "What do you think is the percent chance that [attribute] is harmful to your health?" with the following possible responses: "0–20% (Definitely not harmful)"; "20–40% (Likely not harmful)"; "40–60% (Not sure, either way)"; "60–80% (Likely harmful)"; and "80–100% (Definitely harmful)." As mentioned in our theoretical framework, we treat belief that "[ingredient] is toxic" as a Bernoulli prior; uncertainty is represented by how far the consumer's response is from 0 or 1.[17] Second, we focus attention on the debunking source with the largest effect (magnitude-wise) as revealed in the ingredient studies. For fluoride and aluminum, this source is the regulator, and for GMOs, the competitor. For GMOs, we also include regulator debunking to be consistent with the studies for the other two ingredients. Third, we modify the treatment ad for GMOs. In the GMO ingredient study, our treatment ad stated "certified organic and GMO-free," whereas the control ad does not mention "organic." Thus, the attribute "organic" is confounded with the "non-GMO" attribute. To prevent this confound, we removed the mention of "organic" from the treatment ad.[18] We increased the sample sizes for the belief studies significantly compared with those in the ingredient studies so that there is sufficient statistical power to detect differences across belief groups: 4,758, 5,050, and 4,504 for aluminum, fluoride, and GMOs, respectively. As in the first set of experiments, the sample size for each survey was determined by a pilot study. These surveys are preregistered (https://aspredicted.org/6qn6c.pdf, https://aspredicted.org/6vw3m.pdf, and https://aspredicted.org/ez4cy.pdf for aluminum, fluoride, and GMOs, respectively). Section WB.1 in Web Appendix B provides more details on the belief studies' implementation.

Results of belief studies. Figure W14 in Web Appendix F displays the distribution of participants' prior beliefs across all three ingredients. The average participant ex ante believes that aluminum is harmful, with the most popular prior belief of "likely harmful" (29%). In contrast with

[15] For examples of literature that uses structure to separately identify preferences from beliefs, see Ching, Erdem, and Keane (2013).

[16] Because of repeated measurement 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 answers when asked again, making it seem that they responded to treatment whereas it is in fact a statistical artifact of repeated measurements. We use the control group to control for reversion to the mean.

[17] We note that we cannot ensure that participants' stated beliefs are indeed "true" beliefs. Unlike Ching et al. (2021), who use an incentive-compatible method to elicit beliefs by rewarding accuracy of guesses, our goal is not amenable to this design because we want to measure consumers' subjective beliefs about whether the ingredient is toxic, rather than their best estimate of the scientific consensus assuming that researchers take this as the truth. When their subjective beliefs are different from the scientific consensus on average, rewarding participants on the basis of how close their stated beliefs are to the scientific consensus will bias the measure. To see the differences, consider such an example: a consumer may know that the scientific consensus is that GMOs have no harmful effects on health but still consider GMOs to be harmful to them.

[18] We thank the editor and review team for pointing this out.

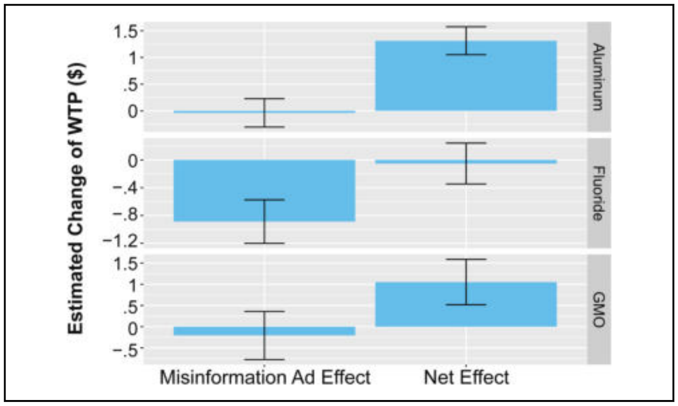

Figure 6. Belief Studies: Average Treatment Effects on WTP.
Notes: The figure displays the average treatment effects on WTP for all three product categories. For ease of comparison with the ingredient studies, we plot the net effect after regulator debunking for all three categories. We report the competitor debunking effect for GMO in Figure W6 in Web Appendix B. Error bars represent the 95% confidence intervals.

aluminum, the vast majority of participants in the fluoride survey believe that fluoride is not harmful: 72% believe that fluoride is not harmful, and only 12% believe that fluoride is harmful. This is consistent with the positive WTP for fluoride measured in the first survey. Beliefs about GMOs are relatively more evenly distributed; the majority (53%) believe that GMOs are not harmful, and 21% believe that GMOs are harmful.

We first estimate the treatment effects averaged over all priors using the same specification as in the previous study (Equation 4) to verify that the average effects replicate. Figure 6 displays the estimated average treatment effects on WTP from the belief studies. Our main takeaways from the ingredient studies replicate: misinformation, on average, decreases WTP for fluoride but has no effect for aluminum and GMOs (same as in Figure 4, Panel B, Column 1). Furthermore, we find that the debunking effects for aluminum and for fluoride also replicate. By increasing the sample size of the GMO ingredient study twofold in the belief study, we are able to detect a significant effect of GMO debunking, by both the regulator and the competitor. The differences in effect sizes for all categories between the two sets of studies are not statistically significant.

Figure W6 in Web Appendix B displays the separate effects of regulator and competitor debunking for GMOs. For the remainder of this section, we omit GMO competitor debunking from the results unless stated otherwise, given that the difference between regulator and competitor GMO debunking is not statistically significant and holding the sources constant across categories allows for better comparability of results.

Next, we measure the belief-specific treatment effects for misinformation and debunking by estimating the following regression:

$$u_{ij} = \sum_{prior}\left(\beta_0^{prior}ing_j + \beta_1^{prior}ing_j I_i^T + \sum_s(\beta_2^{s,prior}ing_j + \beta_3^{s,prior}ing_j I_i^T)I_i^s\right)I_i^{prior} + \alpha price_j + \gamma Z_j + \varepsilon_{ij}. \quad (5)$$

This specification interacts the ingredient dummy (ing_j) and the treatment dummies (I_i^T indicating exposure to misinformation and I_i^s indicating exposure to debunking from source s) with dummies of prior belief groups (I_i^{prior}). Because there are relatively few participants who believe that fluoride and GMOs are "definitely harmful" and that aluminum is "definitely not harmful," we aggregate the priors into three groups: "not harmful" (Prior 1), which consists of respondents with priors "definitely not harmful" and "likely not harmful"; "not sure" (Prior 2); and "harmful" (Prior 3), which consists of respondents with priors "definitely harmful" and "likely harmful." We report the estimates of Equation 5 in Table 9 and plot the treatment effects on WTP in Figure 7, Panel A. The results for the disaggregated priors are reported in Figure W7 in Web Appendix B.

The belief studies' results reveal three main patterns. First, we find that misinformation impacts preferences by creating misbeliefs (Figure 7, Panel A, Column 1). Specifically, consumers who had prior beliefs that aluminum and fluoride are not harmful reduce their WTP for these ingredients after being exposed to misinformation: by $.48 (a 80% reduction from baseline WTP) for aluminum and by $.87 (a 22% reduction from baseline WTP) for fluoride. In addition, misinformation reduces WTP for those who answered "not sure" about fluoride toxicity by $1.13. Unlike the other product categories, misinformation does not significantly influence purchase behavior for any prior belief group for GMOs. We discuss the uniqueness of GMOs further in the next section.

Second, debunking generally increases WTP for those with unsure and misinformed beliefs. Across all three categories, consumers who were a priori unsure about the ingredient increase their WTP by $1–$2 after debunking. For GMOs and aluminum, debunking increases WTP for consumers with "harmful" beliefs by approximately $2, significantly more than the change among consumers with "not harmful" beliefs at the 95% level for aluminum, and at the

Table 9. Belief Studies: Estimates of Equation 5.

	(1) Deodorant	(2) Toothpaste	(3) Nutrition Shake
Ingredient × Prior 1	−.288***	1.482***	−.339***
	(.062)	(.041)	(.050)
Ingredient × Prior 1 × misinformation ad	−.232**	−.324***	−.0735
	(.091)	(.061)	(.072)
Ingredient × Prior 1 × regulator debunking	.590***	.253***	.0513
	(.089)	(.061)	(.0718)
Ingredient × Prior 1 × competitor debunking			.120*
			(.066)
Ingredient × Prior 1 × misinformation ad × regulator debunking	.0809	.0346	.120
	(.127)	(.087)	(.0991)
Ingredient × Prior 1 × misinformation ad × competitor debunking			.0267
			(.098)
Ingredient × Prior 2 × misinformation ad	−1.033***	.390***	−.977***
	(.080)	(.084)	(.089)
Ingredient × Prior 2	.170	−.422***	−.0999
	(.111)	(.132)	(.128)
Ingredient × Prior 2 × regulator debunking	1.017***	.542***	.257**
	(.106)	(.130)	(.119)
Ingredient × Prior 2 × competitor debunking			.113
			(.119)
Ingredient × Prior 2 × misinformation ad × regulator debunking	−.370**	−.0218	.221
	(.148)	(.189)	(.169)
Ingredient × Prior 2 × misinformation ad × competitor debunking			−.0556
			(.179)
Ingredient × Prior 3	−2.142***	−.472***	−1.906***
	(.088)	(.120)	(.138)
Ingredient × Prior 3 × misinformation ad	.211*	−.295	.213
	(.120)	(.185)	(.190)
Ingredient × Prior 3 × regulator debunking	1.072***	.264	.501***
	(.112)	(.191)	(.189)
Ingredient × Prior 3 × competitor debunking			.330*
			(.181)
Ingredient × Prior 3 × misinformation ad × regulator debunking	−.385**	−.0157	−.0694
	(.159)	(.266)	(.257)
Ingredient × Prior 3 × misinformation ad × competitor debunking			−.174
			(.264)
Price	−.479***	−.372***	−.225***
	(.010)	(.009)	(.004)
Balancing attribute	.317***	.855***	.363***
	(.019)	(.021)	(.022)
Brand dummy			
Brand 1	2.725***	.276***	4.054***
	(.042)	(.021)	(.063)
Brand 2	2.528***	.313***	4.142***
	(.042)	(.021)	(.063)
Brand 3	2.042***	−.101***	3.807***
	(.042)	(.018)	(.064)
Brand 4		−.680***	
		(.039)	
Ingredient	Aluminum	Fluoride	GMOs
N	4,758	5,050	4,504

*$p < .1$, **$p < .05$, ***$p < .01$.

Notes: This table presents the treatment effects of misinformation and debunking on consumer preferences. Priors 1, 2, and 3 are "Not harmful," "Not sure," and "Harmful." For the deodorant survey (Column 1), Brands 1, 2, and 3 are Dove, Kopari, and Speed Stick, respectively. For toothpaste (Column 2), Brands 1, 2, 3, and 4 are Colgate, Crest, Risewell, and Tom's of Maine. For nutrition shakes (Column 3), Brands 1, 2, and 3 are Ensure, Orgain, and Soylent. Robust standard errors clustered by individuals are in parentheses.

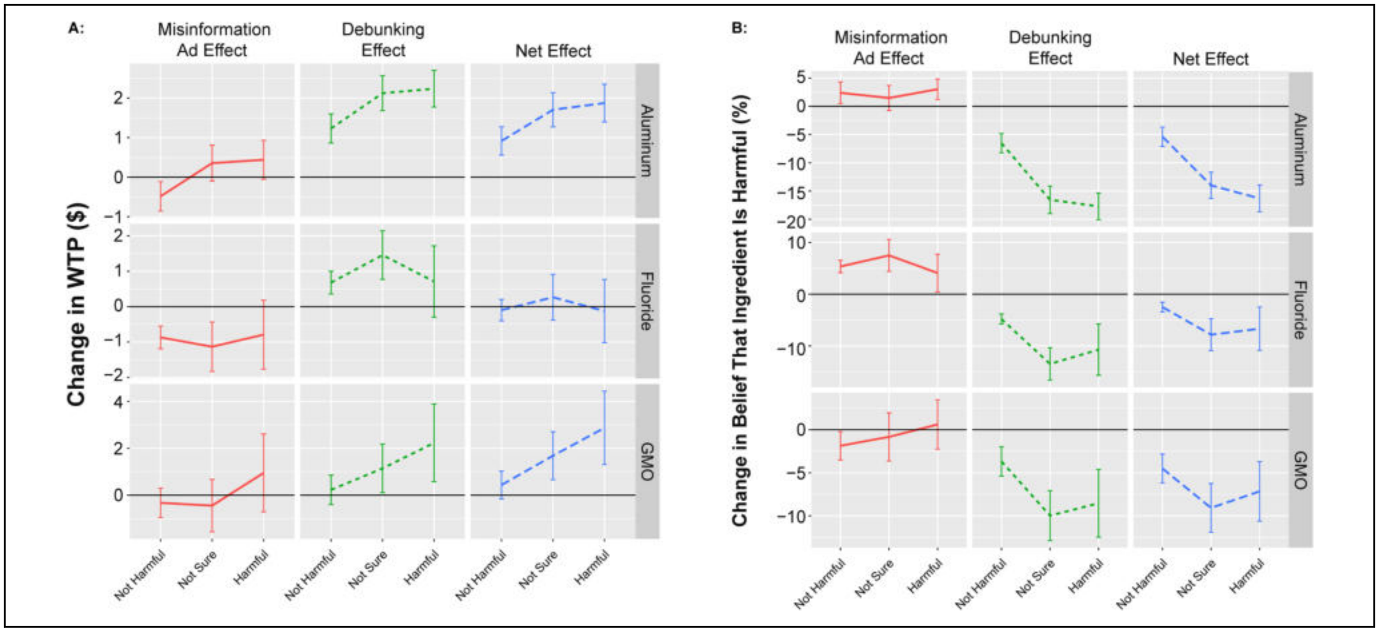

Figure 7. Belief Studies: Treatment Effects.

Notes: This figure plots the estimated treatment effect and 95% CI on WTP (Panel A) and stated beliefs (Panel B) for all three ingredients. The x-axis is the individual's ex ante stated belief. The red line represents comparison across the control group and the misinformation 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 the misinformation ad and debunking. Standard errors are clustered at the user level.

Table 10. Average Misinformation and Net Effects on WTP.

	Misinformation Ad Effect		**Net Effect**	
Study	**Ingredient Study (1)**	**Belief Study (2)**	**Ingredient Study (3)**	**Belief Study (4)**
Aluminum	Nonsignificant	Nonsignificant	Positive	Positive
Fluoride	Negative	Negative	Nonsignificant	Nonsignificant
GMOs	Nonsignificant	Nonsignificant	Nonsignificant	Positive

Notes: "Misinformation ad effect" is the effect of seeing misinformation (comparing participants in the "Misinformation Ad + No Debunk" group with those in the "Control Ad + No Debunking" group). "Net effect" is the effect of seeing an ad with misinformation followed by debunking for participants ("Misinformation Ad + Competitor/Media/Regulator Debunking" compared with "Control Ad + No Debunking").

90% level for GMOs.[19] This suggests that debunking works mainly through correcting misbeliefs instead of reinforcing correct beliefs, and that our debunking treatment can change preferences formed due to *preexisting* misbeliefs. For fluoride, the debunking effect is not statistically significant for those with misinformed beliefs. We note that such insignificance is likely due to the fact that few people have priors that fluoride is harmful: only 12% of participants had the prior belief that fluoride is definitely or likely harmful (Figure W14).

[19] We also find a positive debunking effect for those with "not harmful" priors about aluminum and fluoride; this occurs because the "not harmful" group captures those who are zero to 40% certain that the ingredient is harmful, and thus can still update beliefs.

Third, consistent with the results from the ingredient studies, we find that debunking can revert the reduction in WTP from the experimental misinformation exposure. This is evidenced by the nonnegative net effect (Figure 7, Panel A, Column 3) for the category–prior groups for which misinformation has a negative effect. For these groups, the net effect is statistically greater than the misinformation effect. For instance, fluoride misinformation decreases WTP for fluoride for those with "not harmful" priors, but misinformation *and* debunking does not lead to a net change in WTP. Lastly, we note that the effects on stated beliefs are consistent with the effects on WTP (Figure 7, Panel B).

Overall, we find good news for policy makers: across all categories, the evidence consistently supports the ideal outcome that debunking—and in this case, regulator debunking specifically—works mainly by correcting uncertain and misinformed beliefs.

Table 11. Source-Specific Debunking Effects on WTP.

Source	Debunking Effect Competitor (1)	Media (2)	Regulator (3)
Aluminum	Nonsignificant	Positive	Positive
Fluoride	Positive	Positive	Positive
GMOs	Positive	Nonsignificant	Positive[a]

[a]For GMOs–regulator debunking, we report the effect from the belief study because it uses a much larger sample than the ingredient study. The effect from the ingredient study is statistically insignificant.
Notes: "Debunking effect" is measured by comparing participants in the "Misinformation Ad + Debunking" group with those in the "Misinformation Ad + No Debunking" group.

Discussion

In this section, we summarize the results and provide rationales for differences across sources and product categories. Table 10 demonstrates that the average misinformation and net effects replicate across both studies. Table 11 and Table 12 report the heterogeneous debunking effects by source and by prior beliefs, respectively. The main takeaways from both studies are that (1) misinformation reduces WTP by creating misbeliefs; (2) debunking is effective: debunking reverts the reduction in WTP caused by our experimental dose of misinformation; (3) debunking is effective in correcting misbeliefs: it increases WTP for those with incorrect prior beliefs about the ingredient ("not sure" or "harmful"); and (4) regulator debunking is effective for all categories. Next, we describe in detail the differences across categories and their potential explanations.

Misinformation Effects

Both studies show that the average treatment effect of misinformation is statistically significant and negative for fluoride but insignificant for aluminum and GMOs (Columns 1 and 2 of Table 10). This difference across categories is consistent with —and explained by—the heterogeneous responses to misinformation across prior beliefs. Specifically, the belief studies show that misinformation does not have a significant effect on those who believe the ingredient to be harmful (Column 3 of Table 12). Therefore, in categories where most consumers believe the ingredient to be harmful—that is, aluminum and GMOs (Figure W14)—we observe a null average effect from misinformation.

Another difference across categories is that among participants who believe the ingredient to be "not harmful," misinformation has a significant negative effect for aluminum and fluoride but not for GMOs (Column 1 of Table 12). This could be due to a few reasons. First, the GMO misinformation treatment was the "weakest" of the product categories: the misinformation ad states that the product is GMO-free and thus "has all the good stuff," whereas the misinformation ad for aluminum and fluoride explicitly states that these ingredients are toxic. Second, unlike aluminum or fluoride, GMOs are a unique category for which regulators have implemented policies such as banning and mandatory labeling (see FDA 2023; U.S. Department of Agriculture 2023). Studies have shown that emphasizing GMO disclosure may worsen consumers' perception of GMO safety (Zhang 2014). Such policies from regulators may increase participants' prior exposure to GMO misinformation, making an additional exposure less impactful, regardless of prior beliefs.

Debunking Effects

Effects by source. Although all sources can be effective, the regulator is the only source that is effective across all product categories (Table 11). A potential explanation for the effectiveness of regulator debunking is that regulators are perceived as trustworthy. We find suggestive evidence for this in a follow-up survey among a randomly selected subsample of participants in the fluoride and GMO belief studies who saw the control ad.[20] In the GMO–regulator debunking group, 33% of the participants stated that they found the debunking message to be convincing (i.e., makes them more likely to buy nutrition shakes containing GMOs) because the regulator is trustworthy, whereas the trustworthiness of the firm does not seem to have a strong influence over the participants: only 4% in the firm debunking group found the message convincing *because* it is trustworthy, and only 10% found the message not convincing *because* the firm is not trustworthy.[21]

The preceding statistic invites the question of why competitor debunking is effective for fluoride and GMOs. In the case of fluoride, we conjecture that the source plays a less important role because, unlike in the other categories, the vast majority of consumers already believed that fluoride is not harmful prior to our study: only 12% thought fluoride in toothpaste is harmful, whereas 42% in the aluminum study and 21% in the GMO study considered the ingredient to be harmful (Figure W14). For GMOs, we find suggestive evidence that the novelty of the source predicts the source's effectiveness, in line with Itti and Baldi (2009). Specifically, our GMO follow-up survey reveals that competitor debunking for GMOs is more novel for our study participants than for the other categories: only 15% had seen competitor debunking for GMOs before, compared with 27% ($p < .01$) for fluoride

[20] We selected participants who did not see the misinformation ad because the follow-up survey focuses on debunking. For implementation details of the follow-up survey, see Web Appendix D.

[21] For those that found the debunking message to be convincing, the options for "why" are "The source is trustworthy," "I've heard this message before," "I've always believed that [ingredient X] is safe," "I've tried other [X]-containing products before and they did not cause health issues," or "Other." For those that found the message not to be convincing, the options for "why" are "The source is not trustworthy," "I've always believed that [ingredient X] is safe" (the ceiling effect), "I've heard the opposite from sources," "I need more evidence," "I typically do not buy nutrition shakes/toothpaste," "I believe [X] is not harmful but still prefer to not purchase for other reasons," or "Other."

Table 12. Empirical Effects on WTP by Prior Beliefs.

	Misinformation Ad Effect			Debunking Effect		
Prior	**Not Harmful (1)**	**Not Sure (2)**	**Harmful (3)**	**Not Harmful (4)**	**Not Sure (5)**	**Harmful (6)**
Aluminum	Negative	Nonsignificant	Nonsignificant	Positive	Positive	Positive
Fluoride	Negative	Negative	Nonsignificant	Positive	Positive	Nonsignificant[a]
GMOs	Nonsignificant	Nonsignificant	Nonsignificant	Nonsignificant	Positive	Positive

[a]In the fluoride belief study, few people (12%) reported that they believe fluoride in toothpaste to be harmful. Therefore, the study is underpowered for this belief group.

Notes: This table reports the debunking effect, as measured by comparing participants in the "Control Ad + Debunking" group with participants in the "Control Ad + No Debunking" group.

Table 13. Baseline Equilibrium Market Shares and Prices.

	Market Share (%)		Price ($)	
	Simulated	**Observed**	**Simulated**	**Observed**
Deodorant				
Dove	37.1	46	4.49	4.38
Kopari	1.7	—	13.46	14
Speed Stick	28.4	22	3.97	4.25
Toothpaste				
Colgate	34	33	4.07	3.99
Crest	33.5	34.7	4.2	4.25
Risewell	.5	<5	11.3	12
Tom's of Maine	14.3	—	5.67	5.83
Nutrition Shake				
Ensure	11.5	15	24.59	25
Orgain	37.1	—	23.28	23
Soylent	2.1	—	33.96	34.32

Notes: The table reports the observed and simulated equilibrium market shares and prices for all brands in the baseline market. Observed market shares are obtained from publicly available sources (see Table W22 in Web Appendix C). Public market share data for Kopari, Soylent, Tom's of Maine, and Orgain are unavailable. Observed prices are from Amazon in September 2020.

and 21% ($p = .034$) for aluminum. Further exploring when and why firm debunking works is an interesting avenue for future research.

Effects by prior. We note that the effect of debunking on WTP is generally positive across prior groups in all three categories (Columns 4–6 of Table 12), with two exceptions. First, we do not find a statistically significant effect of debunking among participants who think fluoride is harmful. Once again, this seems to be due to limited sample size in this category–belief combination. Second, for those who believe the ingredient is not harmful, the GMO debunking effect is indistinguishable from zero. This may reflect the combination of uniqueness of GMO (conflicting messages from regulators as discussed previously) and the ceiling effect of debunking for "not harmful" priors. Furthermore, research has shown that GMOs face moral opposition (Scott, Inbar, and Rozin 2016), resulting in largely negative and sustained consumer perception that is difficult to change through scientific reasoning.

Finally, we note that once we doubled the sample size for the GMO survey in the belief study, the average debunking effect became statistically significant and positive (last row of Table 10), again suggesting that the average null effect in GMO debunking in the ingredient study may be due to a lack of statistical power.

Empirical Findings and Theoretical Predictions

We next discuss the interpretation of the empirical findings relative to our theoretical predictions. Recall that the following prediction distinguishes unbiased Bayesian updating from biased updating: unbiased Bayesian updating predicts that those with priors inconsistent with the new information should update their beliefs *more* than those with consistent priors, conditional on the information coming from a trustworthy source. Biased updating predicts the opposite: those with priors inconsistent with the new information should update their beliefs *less* or *in the opposite direction* (i.e., backfire effect) than those with consistent priors.

The following observations demonstrate that our empirical findings are more consistent with unbiased Bayesian updating. First, we find no evidence of the backfire effect; we do not observe debunking to increase misbeliefs or decrease preferences for the focal ingredient for any category or prior. Second, we find that debunking is more effective for those with "harmful" (i.e., inconsistent) priors than with "not harmful" (i.e., consistent) priors for aluminum ($p < .05$), and for GMOs ($p < .1$) (Figure 7, Panel A). Overall, no cases are consistent with biased updating, in which those with "harmful" priors update less (more) than those with "not harmful" priors after debunking (misinformation).

Firm Reactions to Misinformation

Our experiments demonstrate that debunking by competing firms can be effective at changing purchase behavior resulting from misbeliefs. In this section, we investigate whether a firm's best response is to debunk misinformation or to conform to the misinformation by introducing an ingredient-X-free product when faced with an entrant that spreads misinformation in the market. Specifically, we compare the equilibrium outcomes of the incumbents across the following scenarios when an entrant introduces an ingredient-X-free product and spreads misinformation: (1) a focal incumbent debunks while all other incumbents coordinate not to introduce ingredient-X-free products; (2) focal incumbent debunks while other incumbents can add new ingredient-X-free products to their existing portfolio; and (3) no incumbent debunks and all incumbents introduce new ingredient-X-free products. Such comparisons allow us to quantify the trade-offs incumbents face between debunking and complying with misinformed beliefs by introducing an ingredient-X-free product.

We calculate equilibrium market shares and prices given firms' best actions following the Nash equilibrium concept. That is, each firm's equilibrium strategy is the action and price that maximizes its profit, conditional on the other firms' strategies. To compute the equilibrium outcomes, we simulate a market where the products are configured using their Amazon best-seller configurations or their configuration listed on the firm's website (e.g., whitening toothpaste with fluoride).

To simulate a product's market share—and thus profit—we consider a market in which consumers choose between the products on the market and an outside option, which is normalized to have a utility of zero. Given the product's attributes, price, and a product's utility using the parameter estimates from Table 8, we calculate ms_j—product j's market share—using Equation 6 (where the utility follows from Equation 4). We assume that the εs are i.i.d. and have extreme value Type I distribution.

$$ms_j = \frac{\exp(u_{ij})}{1 + \sum_{j'=1,...J} \exp(u_{ij'})}. \tag{6}$$

Because we do not observe marginal cost, we set marginal costs to be those that match equilibrium prices to observed prices. We further assume that the marginal cost for an X-free product is 10% greater than the marginal cost for the X-containing product.

To compute a product's equilibrium price, we use the iterative method described in Allenby et al. (2014): for a given vector of prices, we compute the price that leads to the greatest profit for each product, update the vector of prices with these new prices, and iterate until the price vectors converge.[22] We do this for every action that each incumbent can take to determine the Nash equilibrium strategy.

To describe incumbents' incentives to debunk or to conform to misinformation, we calculate the equilibrium profits for each incumbent under the following scenarios. In all scenarios, the entrant spreads misinformation and offers only an X-free product. All incumbents offer X-containing products to start with.

1. **Baseline**: Each incumbent only offers an X-containing product. There is no debunking.
2. **Coordinated Debunking**: The focal incumbent debunks. Other incumbents do not introduce an X-free product.
3. **Unilateral Debunking**: The focal incumbent debunks. Other incumbents can introduce an X-free product.
4. **No Debunking**: All incumbents each introduce an X-free product.

Comparing Scenario 1 with the baseline allows us to evaluate whether it is profitable for the focal incumbent to debunk under coordinated efforts by other incumbents not to introduce an X-free product. Comparing Scenario 2 with Scenario 1 allows us to evaluate whether other incumbents still have the incentive to conform to misinformation, even in the presence of debunking. Comparing Scenario 3 with the baseline describes whether it is more profitable for all firms to conform to misinformation.

In our simulation, we set the focal incumbent to be the brand that implemented the debunking message in the real world in each product category: Speed Stick, Colgate, and Soylent. We calculate consumers' utilities as though all consumers in the market are exposed to the ads with misinformation and debunking messages, when applicable. Debunking impacts consumers' utilities by changing their preferences for the focal ingredient; therefore, it is not brand-specific.

Table 13 demonstrates that our baseline simulated market shares and prices are similar to the actual U.S. market shares and prices of the brands whose data are publicly available. Figure 8 plots the equilibrium profits for the focal incumbent in all three categories and in all scenarios. Equilibrium market shares, prices, and profits for all firms under each policy can be found in Web Appendix C. We find that debunking increases the focal incumbent's profit relative to the baseline when other incumbents coordinate not to introduce X-free products (Column 2 > Column 1 in Figure 8). This increase in profits comes from the increased preference for ingredient X due to debunking, allowing all incumbents to charge a higher price. However, debunking may hurt the focal incumbent when other incumbents can respond by introducing an X-free product (Column 3 < Column 1 for aluminum and fluoride in Figure 8) and steal the focal incumbent's market share. Expecting such competitor responses, the focal incumbent therefore has the incentive to introduce an X-free product also, resulting in all incumbents conforming to misinformation (Column 4 > Columns 1, 2, and 3 in Figure 8).

For all three categories, our simulations show that all incumbents will choose to introduce an X-free product over debunking when there is no coordinated effort. In the deodorant and nutrition shake categories, firms have a greater incentive to create GMO- and aluminum-free products rather than debunk because consumers are generally averse to these ingredients. For fluoride, the difference in profits for the focal incumbent for debunking compared with creating a fluoride-free product is less stark: \$1.27 versus \$1.28 (Column 2 vs. Column 4), respectively. This

[22] Although this method does not guarantee that the equilibrium is unique, we started the iterative method using several different starting points and were unable to find other equilibria.

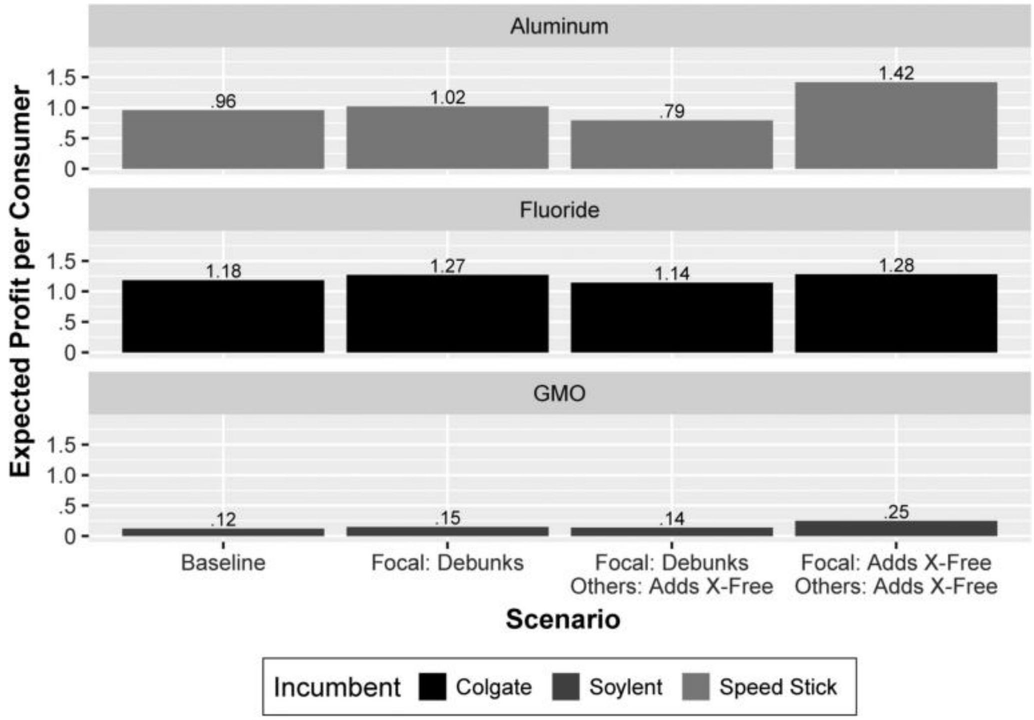

Figure 8. Focal Incumbent's Equilibrium Profit.
Notes: This figure plots the equilibrium profits for the focal incumbent for each product category given the focal incumbent and other incumbents' actions. Equilibrium prices are calculated for all scenarios. The first column is the baseline market in which there is neither debunking nor the introduction of an X-free product by any firm. The second column is the focal firm's profit when the focal firm debunks and other incumbents do not debunk nor introduce an X-free product. In the third column, the focal firm debunks and the other incumbents each introduce an X-free product. In the fourth column, all incumbents each introduce an X-free product. Incumbents are Speed Stick for aluminum, Colgate for fluoride, and Soylent for GMOs.

is because on average, consumers prefer fluoride, even in the presence of misinformation.[23,24]

Overall, the simulation demonstrates that coordination by incumbents not to introduce an X-free product is necessary for debunking to be an incumbent's equilibrium response. Without a coordinated effort, the best response for all incumbents is to conform to misinformation by creating the X-free product. The incentive to conform to misinformation is especially large for incumbents in product categories with ingredients to which consumers are averse. This can explain why firms are commonly observed to add an ingredient-X-free product to their portfolio: Dove and Speed Stick each launched a new line of aluminum-free deodorants; Tom's of Maine, founded in the 1970s, launched fluoride-free toothpastes in 2012; and Ensure launched a GMO-free line in 2019 (see Ensure 2023; Kinonen and Trakoshis 2019; Lady Speed Stick 2023; Tom's of Maine 2012). Our simulations provide an explanation for why so many firms are willing to comply with misinformation but few appear ready to debunk it, especially when the ingredient is already controversial.

Conclusion

This research investigates the extent that debunking via corrective messaging can revert the effects of misinformation, as well as the heterogeneous impacts of misinformation and debunking based on the debunking source and prior beliefs. Through two sets of incentive-compatible survey experiments, we measure these effects on consumers' revealed preferences and stated beliefs. We find that although misinformation can reduce consumers' WTP by creating misbeliefs, debunking provides an effective strategy to correct such influence by reverting consumers' misbeliefs. We also find that debunking can influence consumers' beliefs formed prior to our study.

[23] To ensure the conclusion that adding an X-free product leads to greater profits is not entirely driven by the increase in demand from simply adding a new product, we conduct a robustness check in which the incumbents introduce another X-containing, rather than X-free, product. This check shows that deodorant and nutrition shake firms benefit more from introducing an aluminum- or GMO-free product than from introducing another aluminum- or GMO-containing product. In the case of toothpaste, where consumers prefer fluoride on average, introducing another fluoride-containing product is always the dominant strategy. Nevertheless, for all ingredients, debunking is not the most profitable response by the focal firm.

[24] We note that our calculated profits are based on assumed marginal costs; if production costs for a new product line are high, then it may not be profitable for the firm to conform to misinformation. We tested the sensitivity of our results to different marginal costs, including using the observed price difference between X-containing and X-free products to approximate marginal cost between these versions. Our results are robust.

Although debunking is shown to be effective on average, the heterogeneous impact of debunking is an important consideration in policy evaluation. From a regulator's perspective, debunking should ideally change actions resulting from misinformed beliefs (consistent with unbiased Bayesian updating) rather than merely reinforce those with correct beliefs (consistent with biased updating). 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 uncertain and 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 are effective across all three categories, whereas competitor debunking and media debunking are each effective for two of three categories. Trustworthiness and perceived novelty of the source in the category provide a parsimonious explanation for the differences in debunking source effectiveness across categories. The finding that competitor debunking is effective suggests that markets might be able to self-regulate. However, debunking may not be the incumbent's most profitable equilibrium strategy. Simulations show that introducing a product that conforms to the misinformation leads to a greater increase in per capita profit than debunking. Although debunking by an incumbent is beneficial to all incumbents if they coordinate not to introduce ingredient-X-free products, achieving such coordination in practice might not be feasible. In our context, without a regulator, the socially optimal goal of eliminating misbeliefs would be hard to achieve.

Such dilemmas appear in other contexts as well, with firms debating whether to embrace sustainability goals, engage in corporate social responsibility, or limit greenhouse gas emissions. When these align with consumer preferences, these pledges might be easy to adopt, but when they hurt profits or require coordinated efforts by all firms, regulator intervention might be needed. For misinformation in the marketplace, 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.

This research has some limitations. Although we take the first step to measure the effect of debunking in combating misinformation in ads, our experiments are conducted in a highly controlled setting, in which participants are paying attention to the source and content of the tweets. Our estimation can be interpreted as the best-case scenario for policy makers, which may be used to bound the potential gain regulators can expect from debunking policies. Understanding debunking effectiveness when individuals selectively pay attention is an avenue for future research. Additionally, our focus on using real misinformation and debunking efforts implies that consumers may have preexisting beliefs and attitudes about the sources of these messages. Although this does not impact the measured treatment effects because the existing preferences are held constant across the treatment and control groups, we acknowledge that the treatment effects may be different if the message comes from another source (e.g., if a more popular brand spreads misinformation). Furthermore, by holding constant the number of likes and retweets across treatment conditions, we assume a certain level of acceptance of the ads and debunking messages. Therefore, our experiments cannot speak to how debunking effectiveness changes under other levels of social acceptance; this would be another fruitful area for future research.

Author Note

An earlier version of this article was circulated under the title "Debunking Misinformation in Advertising."

Special Issue Editor

Peter Danaher

Associate Editor

Raghuram Iyengar

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Allenby, Greg M., Jeff D. Brazell, John R. Howell, and Peter E. Rossi (2014), "Economic Valuation of Product Features," *Quantitative Marketing and Economics*, 12 (4), 421–56.

Amazeen, Michelle A., Arunima Krishna, and Rob Eschmann (2022), "Cutting the Bunk: Comparing the Solo and Aggregate Effects of Prebunking and Debunking Covid-19 Vaccine Misinformation," *Science Communication*, 44 (4), 387–417.

Armstrong, Gary M., Metin N. Gurol, and Frederick A. Russ (1983), "A Longitudinal Evaluation of the Listerine Corrective Advertising Campaign," *Journal of Public Policy & Marketing*, 2 (1), 16–28.

Avery, Rosemary J., John H. Cawley, Matthew Eisenberg, and Jonathan 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), 129–39.

Basch, Corey H., Nicole Milano, and Grace C. Hillyer (2019), "An Assessment of Fluoride Related Posts on Instagram," *Health Promotion Perspectives*, 9 (1), 85–88.

Bernhardt, Kenneth L., Thomas C. Kinnear, and Michael B. Mazis (1986), "A Field Study of Corrective Advertising Effectiveness," *Journal of Public Policy & Marketing*, 5 (1), 146–62.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen (2011), "Calorie Posting in Chain Restaurants," *American Economic Journal: Economic Policy*, 3 (1), 91–128.

Burke, Raymond R. and Thomas K. Srull (1988), “Competitive Interference and Consumer Memory for Advertising,” *Journal of Consumer Research*, 15 (1), 55–68.

Centers for Disease Control and Prevention (2020), “Community Water Fluoridation,” (last reviewed January 15, 2020), https://www.cdc.gov/fluoridation/index.html.

Chan, Man-pui Sally, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín (2017), “Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation,” *Psychological Science*, 28 (11), 1531–46.

Charness, Gary and Dan Levin (2005), “When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect,” *American Economic Review*, 95 (4), 1300–1309.

Ching, Andrew T., Tülin Erdem, and Michael P. Keane (2013), “Learning Models: An Assessment of Progress, Challenges, and New Developments,” *Marketing Science*, 32 (6), 913–38.

Ching, Andrew T., Tanjim Hossain, Shervin Shahrokhi Tehrani, and Clarice Yulai Zhao (2021), “How Do People Update Beliefs? Evidence from the Laboratory,” working paper, University of Toronto.

Chiou, Lesley and Catherine Tucker (2018), "Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement,” National Bureau of Economic Research Working Paper 25223, https://doi.org/10.3386/w25223.

Coutts, Alexander (2019), “Good News and Bad News Are Still News: Experimental Evidence on Belief Updating,” *Experimental Economics*, 22 (2), 369–95.

Danaher, Peter J., André Bonfrer, and Sanjay Dhar (2008), “The Effect of Competitive Advertising Interference on Sales for Packaged Goods,” *Journal of Marketing Research*, 45 (2), 211–25.

Ding, Min, Rajdeep Grewal, and John Liechty (2005), “Incentive-Aligned Conjoint Analysis,” *Journal of Marketing Research*, 42 (1), 67–82.

Duggal, Neel (2020), "Aluminum-Free Deodorant: Is It Safe?” Speed Stick (accessed October 2020), https://www.speedstick.com/en-us/sweat/aluminum-free-deodorant-is-it-safe.

Dyer, Robert F. and Philip 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, Robert F. and Phillip G. Kuehl (1978), “A Longitudinal Study of Corrective Advertising,” *Journal of Marketing Research*, 15 (1), 39–48.

Ecker, Ullrich K.H., Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen (2022), “The Psychological Drivers of Misinformation Belief and Its Resistance to Correction,” *Nature Reviews Psychology*, 1, 13–29.

Ensure (2023), “Frequently Asked Questions,” (accessed March 20, 2023), https://ensure.com/nutrition-facts-questions-answers.

Federal Trade Commission (2019), “FTC Obtains Record $191 Million Settlement from University of Phoenix to Resolve FTC Charges It Used Deceptive Advertising to Attract Prospective Students,” press release (December 10), https://www.ftc.gov/news-events/press-releases/2019/12/ftc-obtains-record-191-million-settlement-university-phoenix.

Federal Trade Commission (2020), “Legal Library: Cases and Proceedings,” (accessed September 2020), https://www.ftc.gov/enforcement/cases-proceedings.

Federal Trade Commission (2023), “Truth in Advertising,” (accessed March 2023), https://www.ftc.gov/news-events/topics/truth-advertising.

Food and Drug Administration (2023), “How GMOs Are Regulated in the United States,” (accessed March 20, 2023), https://www.fda.gov/food/agricultural-biotechnology/how-gmos-are-regulated-united-states.

Fryer, Roland G., Jr., Philipp Harms, and Matthew O. Jackson (2019), “Updating Beliefs When Evidence Is Open to Interpretation: Implications for Bias and Polarization,” *Journal of the European Economic Association*, 17 (5), 1470–501.

Funk, Cary, Brian Kennedy, and Meg Hefferon (2018), “Public Perspectives on Food Risks,” Pew Research Center (November 19), https://www.pewresearch.org/science/2018/11/19/public-perspectives-on-food-risks/.

Green, Paul E. and Vithala R. Rao (1971), “Conjoint Measurement for Quantifying Judgmental Data,” *Journal of Marketing Research*, 8 (3), 355–63.

Green, Paul E. and Venkatachary Srinivasan (1978), “Conjoint Analysis in Consumer Research: Issues and Outlook,” *Journal of Consumer Research*, 5 (2), 103–23.

Green, Paul E. and Venkat Srinivasan (1990), “Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice,” *Journal of Marketing*, 54 (4), 3–19.

Guess, Andrew, Brendan Nyhan, and Jason Reifler (2020), “Exposure to Untrustworthy Websites in the 2016 US Election,” *Nature Human Behaviour*, 4 (5), 472–80.

He, Sherry, Brett Hollenbeck, and Davide Proserpio (2022), “The Market for Fake Reviews,” *Marketing Science*, 41 (5), 896–921.

Itti, Laurent and Pierre Baldi (2009), “Bayesian Surprise Attracts Human Attention,” *Vision Research*, 49 (10), 1295–306.

Jindal, Pranav and Anocha Aribarg (2021), “The Importance of Price Beliefs in Consumer Search,” *Journal of Marketing Research*, 58 (2), 321–42.

Keller, Kevin Lane (1991), “Memory and Evaluation Effects in Competitive Advertising Environments,” *Journal of Consumer Research*, 17 (4), 463–76.

Kinonen, Sarah and Angela Trakoshis (2019), “Dove’s First Aluminum-Free Deodorant Is Surprisingly Long-Lasting,” *Allure* (September 25), https://www.allure.com/story/dove-zero-percent-aluminum-deodorant-review.

Kong, Xinyao and Anita Rao (2021), “Do ‘Made in USA’ Claims Matter?” *Marketing Science*, 40 (4), 731–64.

Lady Speed Stick (2023), “Products,” (accessed March 20, 2023), https://www.ladyspeedstick.com/en-us/products.

Lewandowsky, Stephan, John Cook, Klaus Oberauer, Scott Brophy, Elisabeth A. Lloyd, and Michael Marriott (2015), “Recurrent Fury: Conspiratorial Discourse in the Blogosphere Triggered by Research on the Role of Conspiracist Ideation in Climate Denial,” *Journal of Social and Political Psychology*, 3 (1), 161–97.

Lewandowsky, Stephan, Ullrich K.H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook (2012), “Misinformation and Its Correction: Continued Influence and Successful Debiasing,” *Psychological Science in the Public Interest*, 13 (3), 106–31.

Liaukonyte, Jura, Nadia A. Streletskaya, Harry M. Kaiser, and Bradley J. Rickard (2013), “Consumer Response to ‘Contains’ and ‘Free of’

Labeling: Evidence from Lab Experiments," *Applied Economic Perspectives and Policy*, 35 (3), 476–507.

Lorenz-Spreen, Philipp, Michael Geers, Thorsten Pachur, Ralph Hertwig, Stephan Lewandowsky, and Stefan M. Herzog (2020), "Boosting People's Ability to Detect Microtargeted Advertising," PsyArXiv, https://doi.org/10.31234/osf.io/ea28z.

Luca, Michael and Georgios Zervas (2016), "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62 (12), 3412–27.

Manski, Charles F. (2004), "Measuring Expectations," *Econometrica*, 72 (5), 1329–76.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (8), 2421–55.

Mazis, Michael B. and Janice E. Adkinson (1976), "An Experimental Evaluation of a Proposed Corrective Advertising Remedy," *Journal of Marketing Research*, 13 (2), 178–83.

Nickerson, Raymond S. (1998), "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology*, 2 (2), 175–220.

Nyarko, Yaw and Andrew Schotter (2002), "An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica*, 70 (3), 971–1005.

Nyhan, Brendan and Jason Reifler (2010), "When Corrections Fail: The Persistence of Political Misperceptions," *Political Behavior*, 32 (2), 303–30.

Olson, Jerry C. and Philip A. Dover (1978), "Cognitive Effects of Deceptive Advertising," *Journal of Marketing Research*, 15 (1), 29–38.

Palus, Shannon (2019), "Let It All Sweat Out," *Slate* (June 22), https://slate.com/human-interest/2019/06/rise-of-all-natural-deodorants-not-necessary-sweating-antiperspirant.html.

Penn Medicine (2019), "Is Deodorant Harmful for Your Health?" *Health and Wellness* (June 6), https://www.pennmedicine.org/updates/blogs/health-and-wellness/2019/june/deodorant.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David 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*, 66 (11), 4944–57.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand (2020), "Fighting Covid-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention," *Psychological Science*, 31 (7), 770–80.

Pennycook, Gordon and David G. Rand (2019), "Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality," *Proceedings of the National Academy of Sciences*, 116 (7), 2521–26.

Rabin, Matthew and Joel L. Schrag (1999), "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics*, 114 (1), 37–82.

Rao, Anita (2022), "Deceptive Claims Using Fake News Advertising: The Impact on Consumers," *Journal of Marketing Research*, 59 (3), 534–54.

Rao, Anita and Emily Wang (2017), "Demand for 'Healthy' Products: False Claims and FTC Regulation," *Journal of Marketing Research*, 54 (6), 968–89.

Schwartzstein, Joshua (2014), "Selective Attention and Learning," *Journal of the European Economic Association*, 12 (6), 1423–52.

Schwarz, Norbert, Lawrence J. Sanna, Ian Skurnik, and Carolyn Yoon (2007), "Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns," *Advances in Experimental Social Psychology*, 39, 127–61.

Scott, Sydney E., Yoel Inbar, and Paul Rozin (2016), "Evidence for Absolute Moral Opposition to Genetically Modified Food in the United States," *Perspectives on Psychological Science*, 11 (3), 315–24.

Scott, Sydney E. and Paul Rozin (2020), "Actually, Natural Is Neutral," *Nature Human Behaviour*, 4, 989–90.

Tappin, Ben M., Gordon Pennycook, and David G. Rand (2020), "Bayesian or Biased? Analytic Thinking and Political Belief Updating," *Cognition*, 204, 104375.

Tom's of Maine (2012), "New Tom's of Maine SLS and Fluoride-Free Botanically Bright™ Whitening Toothpaste Hits Store Shelves," press release (September 5), https://www.tomsofmaine.com/news/new-toms-of-maine-sls-and-fluoride-free-botanically-bright-whitening-toothpaste-hits-store-shelves.

Tormala, Zakary L. and Richard E. Petty (2004), "Source Credibility and Attitude Certainty: A Metacognitive Analysis of Resistance to Persuasion," *Journal of Consumer Psychology*, 14 (4), 427–42.

Ursu, Raluca, Tülin Erdem, Qingliang Wang, and Qianyun Zhang (2023), "Prior Information and Consumer Search: Evidence from Eye-Tracking," SSRN (February 27), https://doi.org/10.2139/ssrn.4113480.

U.S. Department of Agriculture, Agricultural Marketing Service (2023) "BE Disclosure," (accessed March 20, 2023), https://www.ams.usda.gov/rules-regulations/be.

Walter, Nathan and Sheila T. Murphy (2018), "How to Unring the Bell: A Meta-Analytic Approach to Correction of Misinformation," *Communication Monographs*, 85 (3), 423–41.

Watson, Stephanie (2011), "Antiperspirant Safety: Should You Sweat It?" WebMD (accessed January 2021), https://www.webmd.com/breast-cancer/features/antiperspirant-facts-safety.

Zhang, Juanjuan (2014), "Policy and Inference: The Case of Product Labeling," working paper, Massachusetts Institute of Technology.