universität
wien

Department of Economics

WORKING PAPER

Boris Knapp

Fake Reviews and Naive Consumers

Working Paper No. 2102

July 2021

Fake Reviews and Naive Consumers[*]

Boris Knapp[†]

Abstract

User-generated reviews like those found on Amazon, Yelp, and similar platforms have become an important source of information for most consumers nowadays. It is therefore tempting for firms to manipulate reviews in order to increase demand for their products - but not all consumers are aware of this. We show that in a simple model with fake reviews and naive consumers the unique equilibrium is characterised by partial pooling, where fake reviews blend in with real ones and are persuasive. Policies that reduce the share of naive consumers have opposing effects on the two consumer groups: naives benefit, while sophisticates are harmed. A policy maker concerned with aggregate consumer welfare is thus facing a non-trivial problem. We further show that when real reviews are written strategically, they are not always truthful. Given sufficiently favourable market conditions, the equilibrium where all real reviewers are strategic is outcome equivalent to one where all consumers are sophisticates. In the context of online platforms, where the boundary between consumers and reviewers is fluid, this equivalence result has important practical implications.

JEL Classification: C72; D82; L15

Keywords: Cheap Talk, Naive Receivers, Biased Senders, Fake Reviews

[*] I want to thank my supervisors Maarten Janssen and Philipp Schmidt-Dengler as well as Daniel Garcia, Matan Tsur, Eeva Mauring and Juha Tolvanen for support and helpful advice. This paper has also benefited greatly from a research stay at Kelley School of Business at Indiana University and fruitful conversations with Rick Harbaugh during that time. This research was financially supported by the Austrian Science Fund (FWF) and the Austrian Economic Association (NOeG).

[†] Vienna Graduate School of Economics (VGSE), University of Vienna.
e-mail: boris.knapp@univie.ac.at

1 Introduction

In many markets consumers rely on experts' opinions in order to make educated decisions. In some of these markets, however, they cannot take for granted that experts provide unbiased information. Consider, for example, a platform like Amazon or Yelp, where customers can post reviews that help others make informed choices. Because reviews play an important role in most consumers' purchasing decisions it is tempting for firms to generate favourable reviews that appear to be written by real customers. Various studies, data breaches, and court rulings exposing such cases demonstrate the fact that fake reviews are a severe problem.[1,2,3] For consumers it is often impossible to distinguish between a real review (written by an actual customer) and a fake one (intended to increase demand), so they must account for this unobserved heterogeneity when extracting information from reviews. Some consumers might not even be aware of the fact that reviews are being manipulated and thus trust online reviews to be truthful - they are naive. A recent consumer survey about shopping behaviour[4] finds that 17% fully trust online reviews on Amazon while another 58% somewhat trust them. Focusing on reviews for local businesses on sites such as Google, Facebook, TripAdviser and Yelp, a 2020 survey by Bright Local[5] reports that "79% trust online reviews as much as recommendations from family and friends". Consumer naivety thus seems to be present to a significant extent.

Other markets with similar informational and strategic environments include medical and financial advice. Doctors - especially in the US - often change their prescription behaviour when pharmaceutical companies exert influence. These changes have been linked to adverse health effects, which suggests that the reason for changed prescription behaviour is a distortion of incentives.[6] It is obvious that some patients are oblivious to this fact and generally trust doctors. Even patients who are aware of this, however, generally cannot distinguish a doctor who is affiliated with pharmaceuticals from one who is not. Thus, they cannot know whether advice is coming from a biased or unbiased

[1] Mayzlin, Dover and Chevalier (2014) exploit organizational differences between two platforms to show empirically that reviews differ significantly when faking them is made more difficult. Hu et al. (2012) use a language analysis algorithm to show that across all product categories that they studied fake reviews were present.

[2] In May of 2021, SafetyDetectives research lab (2021) released a report on a breached database containing 13 million direct messages between Amazon vendors and customers willing to provide fake reviews in exchange for free products.

[3] In 2020, the owner of a Korean internet marketing agency was found guilty of writing a total of $35,000$ fake reviews, promoting restaurants in exchange for money, see Ja-young (2021)

[4] CPC Strategy (2019)

[5] Bright Local (2020)

[6] Fernandez and Zejcirovic (2020) demonstrate this in light of the recent opioid crisis in the US.

doctor, but they can take this heterogeneity into account. Credit rating agencies (CRAs) mainly employ two kinds of business models: issuer-pays and investor-pays. When issuers pay for ratings, CRAs might be inclined to issue favourable ratings to please their customers. When investors pay, no such incentives are present. Xia and Strobl (2012) show that ratings are inflated when issuers pay versus when investors pay. Furthermore, the market does not seem to take this difference into account. This suggests that expert heterogeneity and customer naivety might play a role here as well.

This paper studies how naivety is linked to consumer welfare in order to derive interventions that a policy maker might find desirable. Welfare effects of consumer deception is of major interest to most policy makers. For example, *Article 169 of the Treaty on the Functioning of the European Union* defines "consumers' health, safety and economic interests, as well as their right to information" as important objectives. Further directives put emphasis on misleading product information and protecting consumers therefrom.[7] In the US, the Federal Trade Commission is similarly concerned with consumers' welfare and seeks to protect them from misleading marketing.

A simple intervention a policy maker might consider is to inform consumers about the manipulation in the hope of making them less exploitable. We will refer to interventions that decrease the share of naive consumers as *educational policies*. Fake reviews have been covered in different news outlets in recent years.[8] Funding research on and encouraging news coverage of this topic is one possible mechanism to carry out educational policies although other, more direct ones, are conceivable as well.

In this paper we investigate how these educational policies affect consumer welfare. To this end we propose a simple theoretical model that captures the relevant aspects outlined above. It features heterogeneity on the reviewer side, with real reviewers providing helpful information and fake reviewers trying to trick consumers into purchasing. In particular, real reviewers' preferences are *fully aligned* with those of consumers while fake reviewers' preferences are *state independent* or *orthogonal* to those of consumers. In addition, the model accounts for consumer heterogeneity by encompassing sophisticated ones, who understand that some reviews are fake, and naive ones, who do not. More specifically, sophisticated consumers are fully rational Bayesians who hold correct beliefs about the reviewers' equilibrium strategies and thus cannot be systematically deceived while naives take every review at face value.[9]

[7] DIRECTIVE 2005/29/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL as well as DIRECTIVE 2006/114/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL .

[8] see e.g. Pierce et al. (2021) or Bishop (2021)

[9] Deversi, Alessandro and Schwardmann (2020) show that full sophistication and full naivety captures people's behaviour in a signalling game for the most part.

We find that educational policies have three effects. The *direct* effect of moving consumers from the naive to the sophisticated consumer group is always positive because sophisticates enjoy a higher surplus. Additionally, there are two opposing *indirect* effects brought about by a change in the fake reviewer's strategy. To see why, note that, in equilibrium, fake reviewers trade off deceiving sophisticated consumers against deceiving naives. The latter is best achieved by always writing the best possible review but then such a review would not have much credibility with sophisticates as it would almost surely be fake. Writing each review with some probability increases credibility vis-à-vis sophisticates but comes at the cost that low reviews are not effective at tricking naives into buying the product. To resolve this trade-off, fake reviewers mix over all messages above an endogenous threshold, sacrificing persuasiveness towards the naives to some extent but maintaining a certain degree of credibility vis-à-vis the sophisticates. This divides the message space into a separating region below the threshold and a pooling region above it. Messages in the separating region are sent only by real reviewers and are perfectly informative. Those in the pooling region are sent by both reviewer types and are thus compromised. The relative share of the two consumer groups determines how much persuasiveness fake reviewers are willing to give up for additional credibility. Thus, as the share of naives decreases, fake reviewers expand the set of messages they mix over. Because this shrinks the set of perfectly informative messages, sophisticated consumers suffer from reduced *separation.* On the other hand, because fake reviewers shift probability mass from high to moderate reviews, naive consumers are less frequently tricked into purchases. They benefit from decreased *deception.* The overall effect is determined by the relative sizes of these two *indirect* effects as well as the *direct* effect on those consumers who move from the naive to the sophisticate group and therefore enjoy a higher surplus. The example in the following section illustrates these effects.

Furthermore, we study reviewer honesty and its impact on consumer welfare. The benchmark model in Section 3 assumes that all real reviewers are *naively* honest, i.e. they always represent their information truthfully. In contrast to that, we allow real reviewers to be *strategically* honest in Section 6. At first, strategic honesty seems to be a contradiction in itself. What it means, however, is that in his effort to provide helpful information a reviewer might tell "white lies" that benefit consumers. We find that these "white lies" take the form of *underreporting*, i.e. strategically honest reviewers sometimes write reviews that are worse than the actual quality of a product. The reason for this is that in the baseline equilibrium where honest reviews are truthful a high message gets discounted by sophisticated consumers even when it is in fact truthful. Furthermore, the higher the message the more likely it is fake and the more it gets discounted. Therefore, a more moderate review convinces a sophisticate of higher quality. Deviating from truth-telling necessarily harms naives but because consumers benefit in expectation, strategically honest reviewers prefer to underreport for high quality levels.[10] Surprisingly, when real reviewers are all strategically honest, the equilibrium is outcome equivalent to the one where they are naively honest but all consumers are sophisticated. Even though educational policies usually target consumers, my findings suggest that educating reviewers is as effective a policy as well. For review platforms the distinction between consumers and reviewers is not clear-cut because it is consumers who write reviews after all. In this context my findings suggest that consumer education can be implemented as a broad and inclusive policy since it is effective regardless of which side of the market it reaches.

The present work is most closely related to Ottaviani and Squintani (2006), which is the only other paper that studies welfare effects of educational interventions in a sender receiver game.[11] They introduce naive receivers in a cheap talk model based on Crawford and Sobel (1982) and find surprising welfare results: The more likely the receiver is to be naive, the higher her expected welfare. This result follows from the fact that in Crawford and Sobel (1982) the sender's preferences are partially aligned with those of the receiver and therefore if the decision was delegated to the former, the latter would not fare "too bad" despite the bias. Moreover, in Crawford and Sobel (1982) equilibrium communication with a strategic receiver is coarse and the resulting loss of information reduces her welfare. Because the information loss decreases the receiver's welfare more than the bias, it is better for her to be naive (or delegate her decision to the sender) than strategic. In that sense the polar case of Ottaviani and Squintani (2006)'s welfare result is present in Crawford and Sobel (1982). The possibility for the receiver to be of either type with some probability makes equilibrium communication in Ottaviani and Squintani (2006) more complicated. In particular, it is characterised by full separation, i.e. precise but biased communication, for low states of the world and pooling, i.e. coarse but unbiased communication, for high states. Their welfare result, however, is driven by the same fundamental trade-off between information quality (bias) and quantity (coarseness) as in Crawford and Sobel (1982). The higher the probability that the receiver is naive, the more information is transmitted in equilibrium (the larger the set of states for which communication is precise). This increases the welfare of the naive receiver type because biased but precise is better than coarse information. For the

[10] Underreporting is similar in flavour to *reversal* in Smirnov and Starkov (2020) and *countersignalling* in Feltovich, Harbaugh and To (2002) but all three describe distinct phenomena.

[11] Another related paper is Kartik, Ottaviani and Squintani (2007) which differs from Ottaviani and Squintani (2006) only in that it assumes an unbounded state space.

sophisticated receiver type there is not even a trade-off. She prefers biased but precise information because of her ability to debias it. Hence, the welfare of both receiver types increases as the likelihood of naivety goes up and thus Ottaviani and Squintani (2006) conclude that educational policies should be carefully considered due to their perverse effects on consumer welfare. This is in stark contrast to the findings in this paper that educating consumers benefits them in expectation.

Several other papers study communication games with heterogeneity on the sender side (Jindapon and Oyarzun, 2013; Glazer, Herrera and Perry, 2020), on both sides (Chen, 2011), or games where the senders are heterogeneous with respect to their preferences as well as their expertise (Lahr and Winkelmann, 2020). None of these, however, focus on welfare effects of educational policies nor do they study the behaviour of strategic unbiased senders.

An extensive strand of literature has extended the Crawford and Sobel (1982) framework to consider for example multiple receivers (Farrell and Gibbons, 1989), multiple senders (Krishna and Morgan, 2001; Ambrus and Takahashi, 2008), multidimensional state spaces (Battaglini, 2002; Chakraborty et al., 2007), or noisy information (Battaglini, 2004) and has applied it to various settings ranging from issues of political economy (Gilligan and Krehbiel, 1987, 1989; Krishna and Morgan, 2001; Morris, 2001) to stock recommendations (Morgan and Stocken, 2003).

While credible communication is not possible in the Crawford and Sobel (1982) framework when the bias gets too large, Chakraborty and Harbaugh (2010) study communication when the bias is extreme. They show that a multidimensional state space always allows for credible communication, even when the sender's preferences are state independent. Lipnowski and Ravid (2018) also study cheap talk with state independent preferences focusing on the sender's benefits from communication. They assess the value of commitment, thus comparing cheap talk with Bayesian persuasion à la Kamenica and Gentzkow (2011).

We also want to mention recent work on consumer naïveté (Heidhues and Kőszegi, 2010, 2017) and note that it differs from our concept of naivety. While naïveté is about wrong anticipation of own future behaviour, naivety refers to credulity.

The remainder of the paper is structured as follows: An example is provided in the following section. In section 3, we formalise the baseline model where real reviews are always truthful. In section 4, the equilibrium is derived, and in section 5, we analyse comparative statics. We allow for strategic real reviews in section 6 and show that in this extended model telling the truth no longer constitutes an equilibrium strategy. Finally, we conclude in section 7.

2 Example

Consider a consumer who is thinking of buying a certain product and reads a review in order to make an informed decision. The product can be - with equal probabilities - of five quality levels which yield the following utilities: $u(useless) = 0$, $u(bad) = \frac{1}{4}$, $u(average) = \frac{1}{2}$, $u(good) = \frac{3}{4}$ and $u(excellent) = 1$. The expected utility is then $\mathbb{E}[X] = \frac{1}{2}$. With equal probability, the reviewer can be one of two types: he is either real, in which case he is truthful, or fake, in which case his intention is to boost sales. A naive consumer believes the review to be truthful with certainty, while a sophisticated consumer takes into consideration the possibility that the review is fake. Both consumer types purchase the good only if they expect it to exceed their outside option, which can take any value between 0 and 1 and is distributed uniformly.

Let the probability that the consumer is naive be $\frac{1}{2}$ and suppose that fake reviewers always claim the highest quality level. In case the consumer is naive she will believe such a review, think that the product yields a utility of 1, and purchase it with a probability of 1. If the consumer is sophisticated, however, she will take into account that the review might be fake, updating her beliefs according to Bayes' Rule:

$$\begin{aligned} Pr(fake|\text{“}excellent\text{”}) &= \frac{Pr(\text{“}excellent\text{”}|fake) * Pr(fake)}{Pr(\text{“}excellent\text{”}|fake) * Pr(fake) + Pr(\text{“}excellent\text{”}|truthful) * Pr(truthful)} \\ &= \frac{1 * \frac{1}{2}}{1 * \frac{1}{2} + \frac{1}{5} * \frac{1}{2}} = \frac{5}{6} \\ \mathbb{E}^s[X|\text{“}excellent\text{”}] &= Pr(fake|\text{“}excellent\text{”}) * \frac{1}{2} + Pr(truthful|\text{“}excellent\text{”}) * 1 \\ &= \frac{5}{6} * \frac{1}{2} + \frac{1}{6} * 1 = \frac{7}{12} \end{aligned}$$

Hence, upon observing an *“excellent”* review, she will purchase the good with probability $\frac{7}{12}$. The expected purchasing probability that an *“excellent”* review induces is then

$$Pr(purchase|\text{“}excellent\text{”}) = \frac{1}{2} * 1 + \frac{1}{2} * \frac{7}{12} = \frac{19}{24}.$$

Consider now an educational policy that reduces the share of naives from $\frac{1}{2}$ to $\frac{1}{4}$. Suppose fake reviews always claim *“excellent”* quality, as before. The expected purchasing probability would then decrease to $Pr(purchase|\text{“}excellent\text{”}) = \frac{1}{4} * 1 + \frac{3}{4} * \frac{7}{12} = \frac{11}{16} < \frac{3}{4} = Pr(purchase|\text{“}good\text{”})$. This means that deviating to claiming *“good”* quality is more likely to induce a purchase. A fake reviewer must therefore mix over *“good”* and *“excellent”* reviews in such a way that both induce the same purchasing probability.

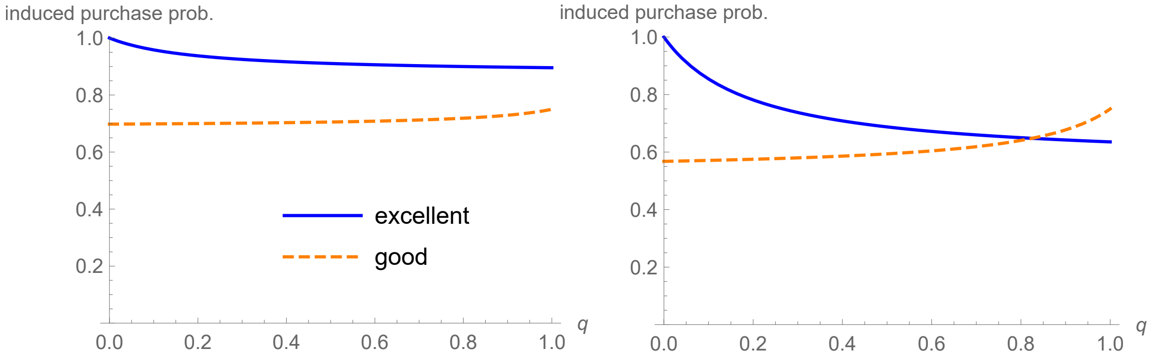

Figure 1: Induced purchasing probabilities as a function of q for $\nu = \frac{1}{2}$ (left) and $\nu = \frac{1}{8}$ (right).

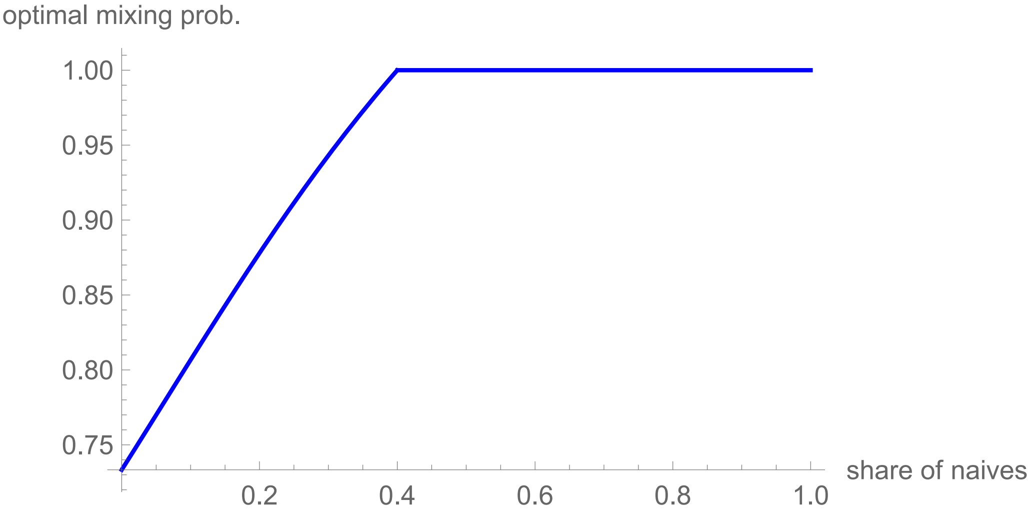

Figure 2: Optimal mixing probability q as a function of the share of naives.

Given that he claims *"excellent"* with prob. q and *"good"* with prob. $1-q$, a sophisticated consumer's posterior beliefs are given by

$$\mathbb{E}^s[X|\,\textit{"excellent"}] = \frac{q * \frac{1}{2}}{q * \frac{1}{2} + \frac{1}{5} * \frac{1}{2}} * \frac{1}{2} + \frac{\frac{1}{5} * \frac{1}{2}}{q * \frac{1}{2} + \frac{1}{5} * \frac{1}{2}} * 1 = \frac{5q+2}{10q+2}$$

$$\mathbb{E}^s[X|\,\textit{"good"}] = \frac{(1-q) * \frac{1}{2}}{(1-q) * \frac{1}{2} + \frac{1}{5} * \frac{1}{2}} * \frac{1}{2} + \frac{\frac{1}{5} * \frac{1}{2}}{(1-q) * \frac{1}{2} + \frac{1}{5} * \frac{1}{2}} * \frac{3}{4} = \frac{13-10q}{24-20q}.$$

and hence the optimal q is given by

$$\frac{1}{4}*1+\frac{3}{4}*\frac{5q+2}{10q+2}=\frac{1}{4}*\frac{3}{4}+\frac{3}{4}*\frac{13-10q}{24-20q} \implies q*=\frac{\sqrt{43}-2}{5}\approx 0.911$$

In Figure 1 the blue and orange lines depict purchasing probabilities induced by *"excellent"* and *"good"* reviews, respectively, when *"excellent"* is claimed with probability q and *"good"* with probability $1-q$, as a function of q. The two panels represent different probabilities that the consumer is naive, $\nu = \frac{3}{4}$ on the left and $\nu = \frac{1}{8}$ on the right. When the likelihood of naivety is sufficiently high, an *"excellent"* review always induces a higher purchasing probability, even when all fake reviews are *"excellent"*. When the consumer is likely enough to be sophisticated, however, claiming *"excellent"* quality exclusively is no longer optimal for a fake reviewer because a *"good"* review then persuades the consumer more often. The point where the curves cross represents the optimal mixing probability q. While in this discrete example mixing kicks in only when the share of naive receivers drops below a certain level, the equilibrium of the continuous model in section 3 will feature mixing for any share of naives (Proposition 1). As the probability of naivety decreases, q goes up, as depicted in Figure 2. Because five star reviews are met with more scepticism by sophisticated receivers, fake senders send more four star reviews as sophisticates become relatively more important (Proposition 2).

On the one hand, this is good news for a naive consumer. Fake reviews that claim *"good"* quality are less deceptive in the sense that the claimed quality is closer to the expected quality.

On the other hand, this is bad for a sophisticated consumer. While she could previously be confident that all reviews except *"excellent"* ones were trustworthy, now *"good"* reviews are compromised as well.

In order to evaluate the effect of this educational policy we examine the ex-ante expected consumer welfare as well as expected consumer welfare conditional on consumer type.[12] To understand the size of the effect despite all normalizations we compare it to the *potential information benefit* (PIB), i.e. the additional surplus a consumer enjoys from having accurate information about the product compared to having no information about the product.

Following the educational policy the expected surplus of a sophisticated consumer falls by around 1.6% of the PIB, while that of a naive consumer increases by around 4%

[12] Expectations are taken over both realisations of product quality and outside option. Ex-ante means that additionally, expectations over consumer type is taken. All calculations can be found in the Appendix.

of the PIB. These are the opposing *indirect* effects that generate the trade-off. Additionally, we need to take into account the *direct* effect of the educational policy which is always positive since sophisticated consumers enjoy greater surplus than naive ones. The overall effect is an increase in ex-ante consumer surplus by roughly 21.4% of the PIB.

In the more general model, we show that in the limiting cases, as the share of naives goes to 0 or 1, educational policies benefit consumers overall (Proposition 3). While we are able to show the positive effect on naives for all parameter values, analytical results for the sophisticates - and therefore also for consumers on aggregate - are provided only in the limiting cases. Numerical calculations suggest that these results hold more generally.

Additionally, we show that if real reviews are written in a strategic way they are not always truthful. Rather, strategically honest reviewers underreport, i.e. they downplay the highest quality levels (Proposition 4). When all real reviewers are strategically honest, the equilibrium outcome is equivalent to the case where all consumers are sophisticated. (Proposition 5)

3 Model

There are two players, a sender and a receiver. Before choosing between a good of unknown quality and her outside option, the receiver reads the sender's review about the product. Throughout the paper we assume that both the quality of the good and the outside option are distributed uniformly on $[0,1]$ and denote the distribution of the outside option Y by F_Y and that of the good X by F_X. While the receiver observes her outside option, she needs to infer the good's quality from the review. A review is a real number $m \in [0,1]$, hence, one can think of them as statements of the form *"The good is of quality m"*.

The sender can be of two types, real or fake, which differ along two dimensions. First, they differ in terms of information. Neither sender type knows the realisation of the receiver's outside option, but while the real type observes the good's quality the fake one does not.[13] Second, they differ with regards to their strategic incentives. The real sender is a non-strategic player who always writes a truthful review, i.e. he passes on his private information honestly.[14] The fake sender is strategic with the objective of

[13] While fake reviewers sometimes do receive the product before writing a review about it, this is not generally the case. We follow the modelling choice of Glazer et al. (2020) to accommodate for both cases. The equilibrium strategy that we derive is optimal also for a fake reviewer who observes the quality.

[14] We will relax this in section 6 and characterise the equilibrium when real senders are strategic and want to maximise the receiver's expected surplus.

inducing a purchase. To this end he is free to send any review $m \in [0,1]$. Formally, we can write his payoff function as $u_S^F = a$ where $a \in \{0,1\}$ denotes the receiver's action of either not buying or buying the good. His preferences are *state independent*, i.e. he cares neither about the good's quality nor the receiver's outside option. Note that a fake sender cannot convey any information with his review since he does not observe the good's quality.

The receiver also can be of two types, sophisticated or naive. A sophisticated receiver is a fully strategic player, who takes into account that the review might be fake, updates her beliefs accordingly, and then takes an optimal action. A naive receiver on the other hand takes every message at face value. The receiver's utility function is $u_R = ax + (1-a)y$, i.e. it is equal to the good's quality x if she makes the purchase $(a = 1)$ and equal to her outside option y if she chooses it instead $(a = 0)$.

Formally, naive and sophisticated receiver types differ in the way they update their beliefs upon seeing a review. A naive receiver updates her beliefs as if every review was truthful. Her expected value of the good's quality, given review m, is

$$\mathbb{E}^n[X|m] = m. \tag{1}$$

Sophisticated receivers take into account that the review might come from a fake sender and not be truthful. By the Law of total expectation, a sophisticate's expected value of the good's quality, conditional on seeing message m, is given by the probability that the sender is fake given that m was sent times the unconditional expected quality (because fake reviews are uninformative), plus the probability that the sender is real given m times m (because real reviews are truthful):

$$\mathbb{E}^s[X|m] = Pr(fake|m)\mathbb{E}[X] + (1 - Pr(fake|m))\, m \tag{2}$$

The receiver buys the product if and only if, after seeing review m, her expected utility of buying the good is above her utility of sticking to her outside option, i.e. if and only if $\mathbb{E}^T[X|m] \geq y$. Because the receiver's choice depends on the review m and the realisation y, we can formulate her choice function as

$$a^T(y,m) = \begin{cases} 1 & \mathbb{E}^T[X|m] \geq y \\ 0 & \mathbb{E}^T[X|m] < y \end{cases}, \tag{3}$$

where $T \in \{s,n\}$ represents the receiver's type. The choice to break indifference in favour of buying is without loss of generality because $\mathbb{E}^T[X|m] = y$ is a probability zero event.

The prior probabilities of the receiver being naive, $\nu \in (0,1)$, and of the sender being fake, $\beta \in (0,1)$, as well as the distributions F_X and F_Y are common knowledge. The timing of the game is as follows:

1. Nature draws x, y, and a type for the receiver and sender.
2. The sender observes his type and - if he is real - also the good's quality. He then sends a review, $m \in [0,1]$, to the receiver.
3. The receiver observes her outside option and the review, and then takes an action $a \in \{0,1\}$.
4. Pay-offs are realised.

As is standard in games with asymmetric information, the solution concept used here is *Perfect Bayesian Equilibrium* (PBE). The fake sender maximises the expected purchasing probability, given his (prior) belief about the receiver's type. A sophisticated receiver maximises her payoff taking the sender's strategy as given and given her beliefs about his type. A naive receiver maximises her payoff taking every review at face value. Formally, a PBE of the game is a pair of purchasing strategies for the two receiver types and a reporting strategy for the fake sender type that fulfill the following conditions:

(a) $a^{s*}(m,y) \in \arg\max_{a\in\{0,1\}} a\mathbb{E}^s[X|m] + (1-a)y$
for all $m \in \mathcal{M}$

(b) $a^{n*}(m,y) \in \arg\max_{a\in\{0,1\}} a\mathbb{E}^n[X|m] + (1-a)y$
for all $m \in \mathcal{M}$

(c) $m_F^* \in \arg\max_{m\in[0,1]} \mathbb{E}_{Y,T}[a^{T*}(m,Y)]$
for all $m^* \in supp(f^F)$

as well as beliefs for the sophisticated receiver type, which are consistent with Bayes' Rule and such that her strategy is optimal given these beliefs.

4 Equilibrium Characterization

We now turn to the equilibrium analysis of the model by looking at the equilibrium conditions at the end of the previous section. Both (a) and (b) require the respective receiver type to maximise expected surplus given her beliefs. According to (c), a biased sender maximises the expected purchasing probability. The subscripts Y and T denote that expectations are taken over both the receiver's type and her outside option. Throughout the paper subscripts denote the variables over which expectations are taken while superscripts are reserved for denoting player types. All proofs can be found in the Appendix while the main text provides the intuition behind the results.

The first result establishes the existence of a unique equilibrium and characterises the fake sender's equilibrium strategy.

Proposition 1 *There exists a unique Perfect Bayesian Equilibrium for any share of fake senders, β, and any share of naive receivers, ν. The reporting strategy of the fake sender is given by $f^F(m) = \frac{1-\beta}{\beta} \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}$ where c is the posterior that fake reviews induce in equilibrium and the unique solution to $\int_c^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m} dm = \frac{\beta}{1-\beta}$.*
Proof. See Appendix.

To better see what the fake sender's equilibrium strategy looks like and understand the intuition behind the equilibrium, it is helpful to look at Figures 3 and 4. Figure 3 depicts the review distributions of the two sender types. Because the honest sender is truthful and the good's quality is distributed uniformly, so are his reviews (blue). The fake sender only sends reviews above a threshold value c (orange). Because he wants to convince the receiver that the quality is high it is intuitive that he would not send low reviews. What is less intuitive is the fact that he mixes over an interval of messages instead of always claiming the highest quality. To understand why, note that an honest sender writes no single review with positive probability. Therefore, were the fake sender to always claim the highest quality in equilibrium, the sophisticated receiver type could infer his type with certainty and ignore the review. Hence, while such a review would be very effective in persuading a naive receiver it would be ineffective in persuading a sophisticated one. A slightly lower review would then not be anticipated by the sophisticate and thus persuade both types. For the same reason the fake sender cannot send any review with positive probability and must mix over a set of messages. The more likely he is to write a certain review the more sceptical a sophisticate is after reading it. In his effort to persuade both types the fake sender puts more probability mass on reviews the higher they are. The increased scepticism of the sophisticated receiver is compensated for by the naive's belief that it is truthful. Figure 4 shows the posterior beliefs of both receiver types together with the expected posterior belief. They differ only for potentially fake reviews. While the naive type's beliefs are equal to the review

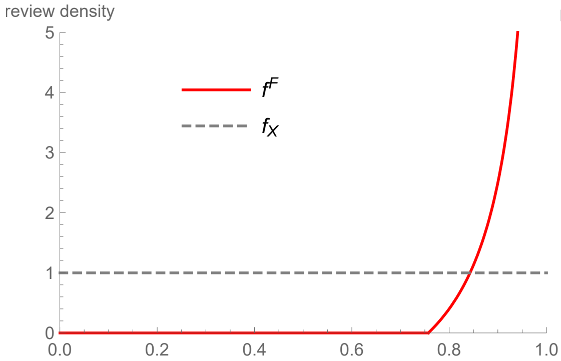

Figure 3: Equilibrium message distributions, $\beta = 0.5$, $\nu = 0.5$

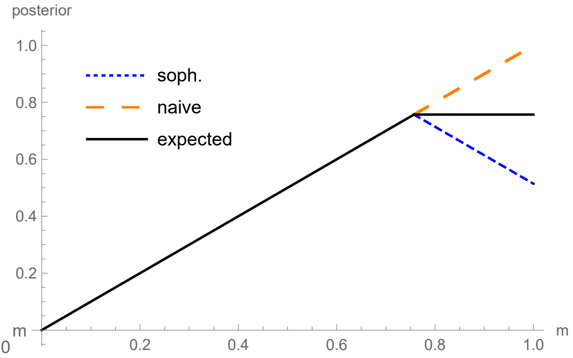

Figure 4: Equilibrium posterior beliefs, $\beta = 0.5$, $\nu = 0.5$

(red), we can see that the sophisticated type grows more sceptical the higher the review (green). This exactly cancels out such that in equilibrium all fake reviews induce the same expected posterior (black).

Put plainly, for a fake review to be effective in equilibrium it needs to blend in with honest ones. Instead of simply claiming that a product is superb a fake review might instead tone it down somewhat and maybe even point out minor flaws, thereby gaining credibility. The model thus predicts that fake reviews do not exclusively praise products, rather, also moderate reviews are potentially fake.

The uniqueness of the equilibrium makes it easy to analyse the comparative statics of the model. In the following section we will study how the equilibrium strategies and receiver surplus change as a result of educational policies.

5 The Effect of Educational Policies

As outlined in the previous section, a fake sender faces a trade-off between deceiving sophisticates, which requires a subtle strategy, and deceiving naives, which is optimally achieved with blatant exaggeration. As we see from Figures 3 and 4, the higher the density that a fake sender puts on a message, the more that message gets discounted by a sophisticated receiver.

Now, if the receiver is more likely to be naive, then the fake sender accepts more scepticism from a sophisticate. Therefore, he increases the density with which he sends high messages. This in turn implies that the set of reviews that he sends in equilibrium shrinks such that the lowest message that he sends is higher. Proposition 2 states that this results in a shift of the fake sender's strategy in the sense of first order stochastic dominance. Moreover, the expected posterior and thus also the purchasing probability increases as the receiver is more likely to be naive.

Proposition 2 *In equilibrium, the posterior induced by a fake sender increases with the share of naive receivers, ν. His reporting strategy shifts mass to high messages in the sense of first order stochastic dominance.*

Proof. see Appendix

At first the two statements in Proposition 2 seem contradictory because an increase in the probability of a given fake review increases the sophisticate's scepticism and therefore *decreases* the posterior. However, it is also more likely that the review is trusted by a naive receiver. The increased scepticism is thus outweighed by the fact that it is less likely to occur. We can see in Figure 5 how the fake reviewer's distribution shifts as the share of naive receiver increases from $\frac{1}{2}$ to $\nu = \frac{3}{4}$.

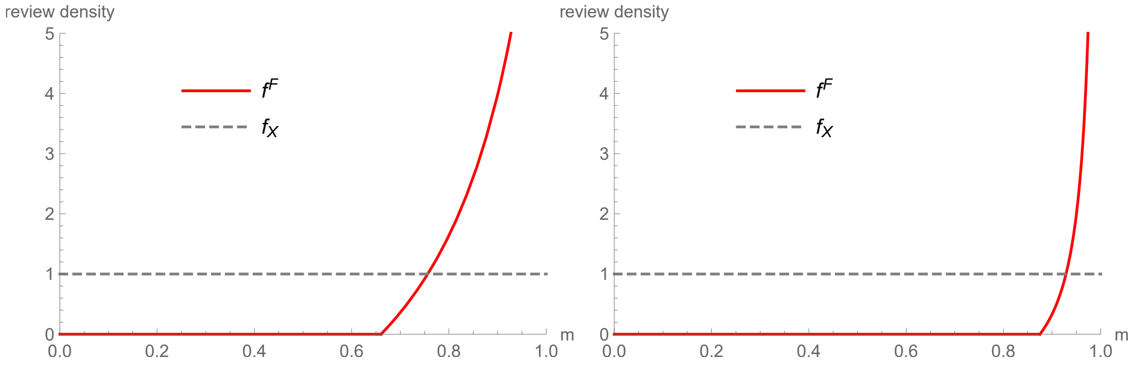

Figure 5: equilibrium message distributions for $\nu = 0.25$ (left) and $\nu = 0.75$ (right)

This paper thus predicts that on platforms with more naive consumers, fake reviews are more blatant. Although the underlying mechanism behind this result is more involved, the idea is very simple. The larger one of two consumer groups gets the more fake senders focus on it. When, on the one hand, the naives become more numerous fake reviews are higher on average because high reviews are effective in persuading naives. On the other hand, as more consumers become sophisticated, fake reviews are lower on average because more subtle reviews persuade sophisticates more effectively.

Consumer Surplus

As pointed out in the introduction, consumer welfare is of particular importance to policy makers. In this paper we therefore focus on how educational policies that change the share of naive consumers affect consumer surplus. In order to evaluate such policies we examine the ex-ante expected consumer surplus. For a naive consumer it is given by:

$$\begin{aligned} CS^n =& \beta \int_c^1 \big(\big[1 - F_Y(m)\big] \, \mathbb{E}[Y|Y > m] + F_Y(m) \, \mathbb{E}[X] \big) \, f^F(m) dm \\ &+ (1-\beta) \int_0^1 \big(\big[1 - F_Y(m)\big] \, \mathbb{E}[Y|Y > m] + F_Y(m) \, m \big) \, dm \end{aligned} \tag{4}$$

Taking a closer look at (4), we can break down CS^n in the following way. Upon observing some review m, a naive consumer either takes her outside option with probability $1 - F_Y(m)$, i.e. if it is greater than m, or buys the good with probability $F_Y(m)$, i.e. if her outside option is below m. In case she takes her outside option, her expected payoff is the expectation of Y, conditional on it being above m. If she buys the good her payoff is m in case the review was real, while her expected payoff is $\mathbb{E}[X]$ in case it was fake. With probability β the review is fake and sent with density $f^F(m)$, while with the remaining probability it is real and sent with uniform density.

A sophisticated consumer's expected surplus is computed equivalently, but taking into account that she forms correct posterior beliefs and buys only if y is above her posterior expectation.

$$\begin{aligned} CS^s =& \beta \int_c^1 \big(\big[1 - F_Y(\mathbb{E}^s[X|m])\big] \, \mathbb{E}[Y|Y > \mathbb{E}^s[X|m]] + F_Y(\mathbb{E}^s[X|m]) \, \mathbb{E}[X] \big) \, f^F(m) dm \\ &+ (1-\beta) \int_0^1 \big(\big[1 - F_Y(\mathbb{E}^s[X|m])\big] \, \mathbb{E}[Y|Y > \mathbb{E}^s[X|m]] + F_Y(\mathbb{E}^s[X|m]) \, m \big) \, dm \end{aligned} \tag{5}$$

Aggregate consumer surplus is then simply the weighted average of CS^n and CS^s:

$$CS = \nu \, CS^n + (1-\nu) \, CS^s. \tag{6}$$

In order to assess the effect of educational policies we have to evaluate the marginal effect of ν, the likelihood of naivety, on consumer surplus:

$$\frac{dCS}{d\nu} = (CS^n - CS^s) + \nu \frac{dCS^n}{d\nu} + (1-\nu) \frac{dCS^s}{d\nu}. \tag{7}$$

An *increase* in likelihood of naivety - the *opposite* of an educational policy - affects aggregate consumer surplus in three ways. The first term on the right hand side of equation 7 captures the effect of moving consumers from the sophisticated to the naive group, this is the *direct effect* of (un-)educating consumers. The second and third term correspond to the *indirect effects* on the naive and sophisticated consumver group, respectively.

The first term is always (weakly) negative. Were a naive consumer to enjoy a higher surplus, a sophisticate could simply imitate her since she has access to at least as much information. In some game-theoretical models it can be beneficial for a player to have access to *less* information. This, however, is due to equilibrium effects, because players may condition their strategies on the information available to the other player. In this model, the consumer's type is private information and therefore imitating a naive consumer's strategy cannot lead to an increase in surplus via equilibrium effects because the sender's strategy wouldn't change.

To understand the effect on the naives, one must bear in mind that every fake review is deceptive in the sense that it claims a quality above the unconditional expectation even though it is sent independently of x. When consumers are more likely to be naive, fake senders shift probability mass to higher reviews (Proposition 2), and hence deceive naive consumers more severely. This affects their surplus negatively. Because this is due to increased deception we call the effect on the naives *deception effect*.

The effect on the sophisticates can intuitively be understood as follows: When a fake sender uses a more blatant strategy as a response to an increase in ν, his messages are stronger signals about his type. Moreover, because the lowest message sent by a fake sender increases, there is a larger set of messages that reveal the good's quality. Loosely speaking, we can say that separation increases and we therefore term the effect on sophisticates *separation effect*. A sophisticate benefits from increased separation so this effect is positive. Analytically, we are able to show this for the limiting cases as $\nu \to 0$ and $\nu \to 1$. Numerical calculations depicted in Figure 6 suggest that this result holds for any share of naive consumers.

Because the deception effect and the separation effect oppose each other, maximising aggregate consumer surplus becomes a non-trivial problem. Increasing the welfare of one of the groups comes at the cost of harming the other. A non-discriminatory policy maker should naturally consider the effect on aggregate consumer welfare. As we see from Figure 6, educational policies benefit consumers on aggregate. Analytically, we are again able to show this for the limiting cases. Proposition 3 summarises these results.

Proposition 3 *In the unique equilibrium, as the share of naive consumers, ν, increases,*

(i) naive consumers are harmed,

(ii) sophisticated consumers benefit in the limit as $\nu \to 0$ and $\nu \to 1$,

(iii) aggregate consumer surplus decreases in the limit as $\nu \to 0$ and $\nu \to 1$.

Proof. see Appendix

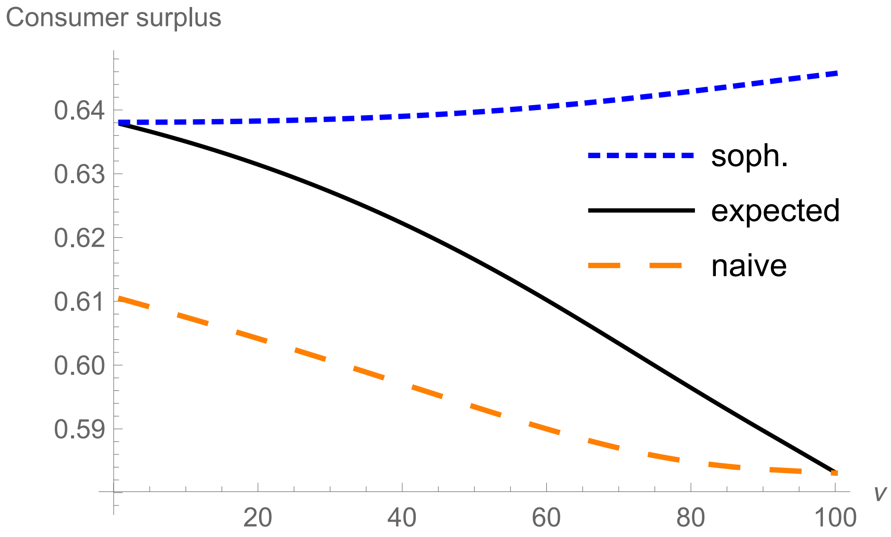

Figure 6: naive, sophisticated, and aggregate consumer surplus as a function of the share of naive consumers

Figure 6 shows how consumer surplus for naives (top left), sophisticates (top right), and aggregate receiver surplus (bottom) change as a function of ν. On the x-axis is the share of naive consumers, from 0% to 100%, and on the y-axis the consumer surplus of the respective consumer group and for the aggregate. Right away we can see that the direct effect dominates as the change in aggregate consumer surplus is larger than that in any of the individual consumer groups. In particular, because the welfare gap between the two groups widens as ν increases, the direct effect becomes larger and larger. Therefore, educational policies are most effective when naivety is very prevalent in a market. Maybe more importantly, the surplus of naive consumers goes up as a result of educational policies. Bear in mind that those are the consumers who were not reached by the policy and are *still* naive. Hence, even the "most vulnerable" - those who are difficult to educate - are reached indirectly with educational policies.

6 Strategic Honesty

In the baseline model we assumed that real reviewers were behavioral types, who always told the truth. One could, however, argue that even real reviewers might misrepresent their information if this mitigates the negative effect of fake reviews. They might behave strategically rather than behaviorally honestly.

In this section we let a fraction η of the real senders be behavioral types just as in the baseline model. We allow the remaining fraction $1-\eta$ to play strategically, i.e. to choose any message in order to maximize the consumer's expected surplus. A strategically honest reviewer's payoff is equal to the receiver's expected utility capturing the idea that people write online reviews because they want to help others make good decisions.

Formally, we augment the baseline game with an additional equilibrium condition for the strategically honest sender:

(c) $m_n^* \in \arg\max\limits_{m\in[0,1]} \mathbb{E}_{Y,T}[a^{T*}(m,Y)x + (1-a_T^*(m,Y))Y]$
for all $m^* \in supp(F^H_{m|x})$ and $x \in [0,1]$

In contrast to Jindapon and Oyarzun (2013) and Glazer et al. (2020) full honesty does not arise in equilibrium. Instead, a strategically honest reviewer engages in *underreporting.*

Proposition 4 *A Perfect Bayesian Equilibrium, where all real reviewers tell the truth, does not exist.*
Proof. See Appendix.

The intuition behind this result is captured by Figure 8, which shows the posterior expectations of the two consumer types given reviewer strategies according to the equilibrium of the baseline model. While truthful reviews are optimal for naive consumers, they are met with scepticism by sophisticates even when they are not fake. Therefore, a high truthful review is not very helpful for them as they will sometimes choose their outside option even though purchasing the good would have been the better choice. Sending a review that is met with less scepticism helps sophisticates, who make mistakes less often as a result. Figure 8 illustrates such a deviation from a truthful review m' to a strategic one m''. Whenever the sophisticated consumer's outside option y is between the true quality m' and her posterior $\mathbb{E}^s[X|m'] < m'$, she forgoes the additional utility $m'-y$. Thus, the green area (light + dark) represents the mistake she *avoids* making in

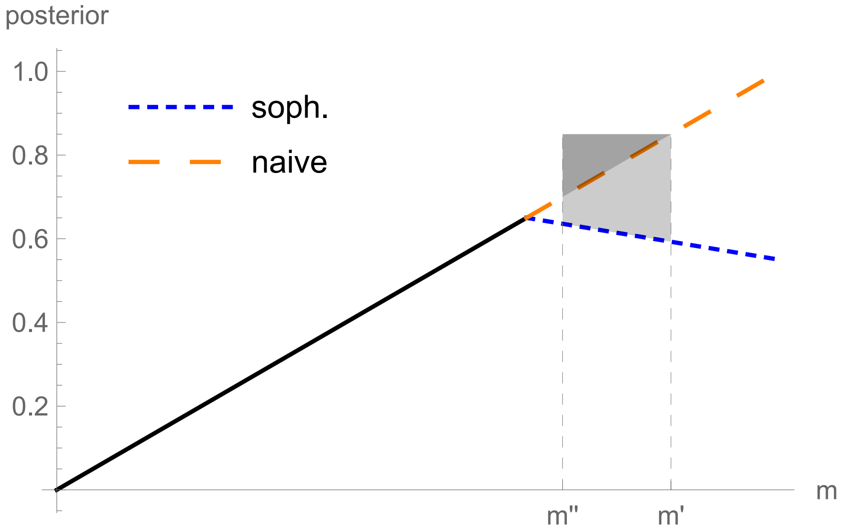

Figure 7: Deviation by a strategically benevolent expert

expectation. Similarly, the dark green area represents the mistake that a naive consumer makes in expectation as a result of the strategic deviation. The light green area then illustrates the net benefit of such a strategic deviation. Proposition 5 characterizes the equilibrium of the extended game.

Proposition 5 *Suppose that a real reviewer is strategically honest with strictly positive probability* $1 - \eta$*. A Perfect Bayesian Equilibrium exists and is unique. A strategically honest reviewer reports truthfully for* $x < c$ *and sends* $m = c$ *whenever* $x \geq c$*. A fake reviewer mixes over the interval* $[c, 1]$*. He sends* $m = c$ *with probability* δ *and messages in* $(c, 1]$ *according to density*

$$f^F(m) = (1 - \delta)\frac{1 - \beta}{\beta}\frac{m - c}{c - (1 - \nu)\mathbb{E}[X] - \nu m} \tag{8}$$

where $\delta = (1 - \eta)\,\frac{1-\beta}{\beta}\frac{(1-c)^2}{2c-1}$ *and* c *is given by the solution to*

$$\int_c^1 \frac{m - c}{c - (1 - \nu)\mathbb{E}[X] - \nu m} dm = \frac{\beta}{1 - \beta}\frac{1 - \delta}{\eta} \tag{9}$$

Proof. See Appendix.

In this equilibrium, a strategically honest reviewer is *partially* truthful but he underreports when the quality is very high because sending a lower review seems more credible.

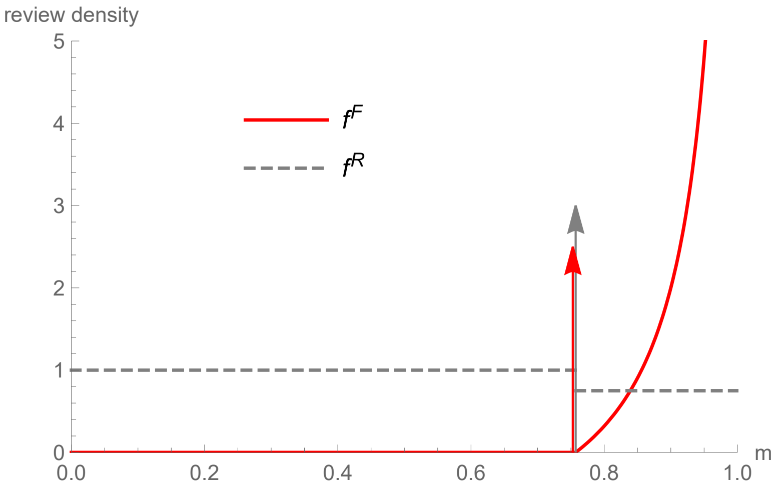

Figure 8: Density of real (gray, dashed) and fake (red) reviews in an equilibrium with strategically honest reviewers.

A fake sender then feeds off of this credibility and also sends lower reviews more often.

In this sense, underreporting by strategically honest reviewers induces fake ones to send more moderate reviews. By an argument similar to that in section 5, naive consumers benefit from fake reviews being less deceptive. In fact, the analogy to educational policies is not too far off as, in the limit, both - the reduction of naive consumers and of behaviorally honest reviewers - can result in the same market outcome.

Proposition 6 *Let $\nu < \frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}$. As the share of behaviorally honest senders, η, disappears, the equilibrium outcome is equivalent to the case where the share of naives, ν, goes to zero.*

Proof. See Appendix.

The intuition behind this result is the following. As η vanishes, no real review is higher than c because all real reviewers underreport. Therefore, fake reviewers also avoid sending messages above c. Those high messages, however, were exactly those that deceived naive consumers. As a consequence, a naive consumer is as well off as a sophisticated one. Under unfavourable market conditions - when the shares of naive consumers *and* that of fake reviewers are high - this equilibrium breaks down.[15]

[15] To see why this equilibrium breaks down if $\nu > \frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}$, suppose $\nu \to 1$. Then sending $m = 1$ induces a posterior arbitrarily close to 1 and thus constitutes a profitable deviation for a fake reviewer in any equilibrium where he induces a posterior below 1. It can be shown that the equilibrium in this parameter region is characterized by fake reviewers mixing between $m = c^{JO}$ and $m = 1$.

The implication of this result is that educational policies can be effective even when they fail to reach consumers if they instead make reviewers strategic. In a context where the distinction between consumers and reviewers on review platforms is fluid, this result is particularly compelling. A way to think about Proposition 6 is the following: Consumers differ in their responsiveness to educational policies and some might be very difficult to reach. These "most naive" consumers are not likely to consume a lot of media or interact much with the world around them. It is not far-fetched to assume that such consumers are also less likely to write reviews. Contrarily, consumers who actively engage on review platforms are likely easier to reach with educational policies. Proposition then states that the fully sophisticated outcome can be obtained even when not all consumers are susceptible to educational policies as long as those who write reviews are.

7 Conclusion

This paper shows that in a market with fake reviews and naive consumers educational policies trade off the surplus of naive consumers against that of sophisticated ones. Because the positive effects outweigh the negative, aggregate consumer surplus increases following educational policies. Contrary to models without naive consumers, writing honest reviews is not always optimal for reviewers who want to maximise consumer surplus. Instead, underreporting arises in equilibrium, such that they downplay a good's quality if it is very high. If all reviews are written strategically, the outcome is equivalent to one where all consumers are sophisticated, given that market conditions are not too unfavourable.

This finding has interesting practical implications. Online platforms do not have a clear boundary between reviewers and consumers. Educational policies that target reviewers will therefore also reach consumers and vice versa. This paper finds that in the limit the outcome is equivalent, regardless of whether all consumers, all reviewers, or both, are educated. Therefore, such policies need not necessarily have a narrow focus but can be implemented broadly.

One may think that the focus on two types of consumers is restrictive, however, in a recent experimental study, Deversi et al. (2020) find that the dichotomy of naivety and full sophistication, for the most part, captured subjects' behaviour in a sender-receiver game. We therefore believe that our modelling choice is well suited to capture consumer heterogeneity on online review platforms and similar markets.

The model in this paper analyses the interaction between a single sender and a single receiver. While accounting for more than one receiver is straightforward,[16] the model does not easily generalise to multiple senders. Glazer et al. (2020) analyse a model with multiple - fake or honest - reviewers and multiple sophisticated consumers and study learning dynamics. They show that in the limit, as a consumer reads an increasing number of reviews, she learns the true quality almost surely as long as there is a positive measure of honest reviews. Their approach hinges on an independence result which does not continue to hold in the presence of naive consumers.[17] In such a model, the equilibrium outcome is likely to depend on how naive consumers are modelled to interpret multiple, conflicting messages. Psychological phenomena like confirmation bias might also play a crucial role: Out of multiple reviews, a consumer may overweigh the ones that confirm her prior. This is an interesting avenue for further research.

[16] For example, if a review is read by a mass of receivers with a share ν of naives, the sender's utility is equal to the expected utility in this model.

[17] They show that the fake reviewers strategy is independent of the history of reviews and are thus able to derive the equilibrium of the multi-sender game.

Bibliography

Ambrus, Attila and Satoru Takahashi, "Multi-sender cheap talk with restricted state spaces," *Theoretical Economics*, 2008, *3* (1), 1–27.

Battaglini, Marco, "Multiple referrals and multidimensional cheap talk," *Econometrica*, 2002, *70* (4), 1379–1401.

_ , "Policy advice with imperfectly informed experts," *Advances in theoretical Economics*, 2004, *4* (1).

Bishop, Todd, "The secrets of Amazon reviews: Feedback, fakes, and the unwritten rules of online commerce," `https://web.archive.org/web/20210526230044/https://www.geekwire.com/2021/secrets-amazon-reviews-feedback-fakes-unwritten-rules-online-commerce/` 2021. accessed: 2021-05-31.

Bright Local, "Local Consumer Review Survey 2020," `https://www.brightlocal.com/research/local-consumer-review-survey/` 2020. accessed: 2021-05-31.

Chakraborty, Archishman and Rick Harbaugh, "Persuasion by cheap talk," *American Economic Review*, 2010, *100* (5), 2361–82.

_ , _ et al., "Comparative cheap talk," *Journal of Economic Theory*, 2007, *132* (1), 70–94.

Chen, Ying, "Perturbed communication games with honest senders and naive receivers," *Journal of Economic Theory*, 2011, *146* (2), 401–424.

Council of European Union, "DIRECTIVE 2005/29/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council," *Official Journal of the European Union*, 2005-06-11, *L 149/22*, 22–39.

_ , "DIRECTIVE 2006/114/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 12 December 2006 concerning misleading and comparative advertising," *Official Journal of the European Union*, 2006-12-27, *L 376/21*, 21–27.

_ , "Article 169 (ex Article 153 TEC)," *Official Journal of the European Union*, 2016-06-07, *C 202/124*, 124–124.

CPC Strategy, "The 2019 Amazon Consumer Shopping Study," `https://tinuiti.com/content/guides/2019-amazon-consumer-shopping-study/` 2019. accessed: 2021-05-31.

Crawford, Vincent P and Joel Sobel, "Strategic information transmission," *Econometrica: Journal of the Econometric Society*, 1982, pp. 1431–1451.

Deversi, Marvin, Ispano Alessandro, and Peter Schwardmann, "Spin Doctors: An Experiment on Vague Disclosure," *working paper*, 2020.

Farrell, Joseph and Robert Gibbons, "Cheap talk with two audiences," *The American Economic Review*, 1989, *79* (5), 1214–1223.

Feltovich, Nick, Richmond Harbaugh, and Ted To, "Too cool for school? Signalling and countersignalling," *RAND Journal of Economics*, 2002, pp. 630–649.

Fernandez, Fernando and Dijana Zejcirovic, "The Role of Pharmaceutical Promotion to Physicians in the Opioid Epidemic," *working paper*, 2020.

Gilligan, Thomas W and Keith Krehbiel, "Collective decisionmaking and standing committees: An informational rationale for restrictive amendment procedures," *Journal of Law, Economics, & Organization*, 1987, *3* (2), 287–335.

_ and _ , "Asymmetric information and legislative rules with a heterogeneous committee," *American Journal of Political Science*, 1989, pp. 459–490.

Glazer, Jacob, Helios Herrera, and Motty Perry, "Fake Reviews," *Economic Journal (forthcoming)*, 2020.

Heidhues, Paul and Botond Kőszegi, "Exploiting naivete about self-control in the credit market," *American Economic Review*, 2010, *100* (5), 2279–2303.

_ and _ , "Naivete-based discrimination," *The Quarterly Journal of Economics*, 2017, *132* (2), 1019–1054.

Hu, Nan, Indranil Bose, Noi Sian Koh, and Ling Liu, "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," *Decision support systems*, 2012, *52* (3), 674–684.

Ja-young, Yoon, "Man jailed for fake restaurant reviews," `https://web.archive.org/web/20210526154538/http://www.koreatimes.co.kr/www/nation/2021/05/251_309423.html` 2021. accessed: 2021-05-31.

Jindapon, Paan and Carlos Oyarzun, "Persuasive communication when the sender's incentives are uncertain," *Journal of Economic Behavior & Organization*, 2013, *95*, 111–125.

Kamenica, Emir and Matthew Gentzkow, "Bayesian persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Kartik, Navin, Marco Ottaviani, and Francesco Squintani, "Credulity, lies, and costly talk," *Journal of Economic Theory*, 2007, *134* (1), 93–116.

Krishna, Vijay and John Morgan, "A model of expertise," *The Quarterly Journal of Economics*, 2001, *116* (2), 747–775.

Lahr, Patrick and Justus Winkelmann, "Fake Experts," *Discussion Paper Series - CRC TR 224, Discussion Paper No. 093*, 2020.

Lipnowski, Elliot and Doron Ravid, "Cheap talk with transparent motives," *Available at SSRN 2941601*, 2018.

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

Morgan, John and Phillip C Stocken, "An analysis of stock recommendations," *RAND Journal of economics*, 2003, pp. 183–203.

Morris, Stephen, "Political correctness," *Journal of political Economy*, 2001, *109* (2), 231–265.

Ottaviani, Marco and Francesco Squintani, "Naive audience and communication bias," *International Journal of Game Theory*, 2006, *1* (35), 129–150.

Pierce, Matthew, Jonathon Gatehouse, Alex Shprinsten, and Meara Belanger, "Black market in Google reviews means you can't believe everything you read," `https://web.archive.org/web/20210526144036/https://www.cbc.ca/news/investigates/fake-reviews-on-google-1.6033859` 2021. accessed: 2021-05-31.

SafetyDetectives research lab, "Amazon Fake Reviews Scam Exposed in Data Breach," `https://web.archive.org/web/20210525123756/https://www.safetydetectives.com/blog/amazon-reviews-leak-report/` 2021. accessed: 2021-05-31.

Smirnov, Aleksei and Egor Starkov, "Bad News Turned Good: Reversal Under Censorship," *AEJ: Microeconomics (forthcoming)*, 2020.

Xia, Han and Gunter Strobl, "The Issuer-pays Rating Model and Rating Inflation: Evidence from Corporate Credit Ratings," *working paper, unpublished*, 2012.

Appendix

Example. Given that fake reviews are *"excellent"* with prob. q and *"good"* with prob. $1-q$, the expected product qualities conditional on these claims, according to Bayes' Rule, are

$$\begin{aligned}
\mathbb{E}^s[X \textit{``excellent''}] &= Pr(fake|\textit{``excellent''}) * \frac{1}{2} + Pr(truthful|\textit{``excellent''}) * 1 \\
&= \frac{q * \frac{1}{2}}{q * \frac{1}{2} + \frac{1}{4} * \frac{1}{2}} * \frac{1}{2} + \frac{\frac{1}{4} * \frac{1}{2}}{q * \frac{1}{2} + \frac{1}{4} * \frac{1}{2}} * 1 = \frac{2q+1}{4q+1} \\
\mathbb{E}^s[X \textit{``good''}] &= Pr(fake|\textit{``good''}) * \frac{1}{2} + Pr(truthful|\textit{``good''}) * \frac{3}{4} = \frac{11-8q}{20-16q}.
\end{aligned}$$

The optimization problem for a fake reviewer is to maximise purchasing probability and boils down to choosing the mixing probability $q = min\{q*, 1\}$, where $q*$ equalises the purchasing probability induced by claiming *"excellent"* with probability q and claiming *"good"* with probability $(1-q)$.

Let Y be the outside option of a consumer and $Y \sim U[0,1]$. A sophisticated consumer's expected surplus can be written as

$$\begin{aligned}
\mathbb{E}CS^s(q) = \frac{1}{2}\Bigg(q\Bigg[&Pr(Y \leq \mathbb{E}^s[X\textit{``excellent''}])\frac{1}{2} + Pr(Y > \mathbb{E}^s[X|\textit{``excellent''}])\mathbb{E}[Y|Y > \mathbb{E}^s[X|\textit{``excellent''}]]\Bigg] \\
&+ (1-q)\Bigg[Pr(Y \leq \mathbb{E}^s[X|\textit{``good''}])\frac{1}{2} + Pr(Y > \mathbb{E}^s[X|\textit{``good''}])\mathbb{E}[Y|Y > \mathbb{E}^s[X|\textit{``good''}]]\Bigg]\Bigg) \\
&+ \frac{1}{2}\left(\frac{1}{5}\frac{1}{2} + \frac{1}{5}\left[Pr(Y \leq \frac{1}{4})\frac{1}{4} + Pr(Y > \frac{1}{4})\mathbb{E}[Y|Y > \frac{1}{4}]\right]\right. \\
&+ \frac{1}{5}\left[Pr(Y \leq \frac{1}{2})\frac{1}{2} + Pr(Y > \frac{1}{2})\mathbb{E}[Y|Y > \frac{1}{2}]\right] \\
&+ \frac{1}{5}\left[Pr(Y \leq \mathbb{E}^s[X|\textit{``good''}])\frac{3}{4} + Pr(Y > \mathbb{E}^s[X|\textit{``good''}])\mathbb{E}[Y|Y > \mathbb{E}^s[X|\textit{``good''}]]\right] \\
&\left.+ \frac{1}{5}\Big[Pr(Y \leq \mathbb{E}^s[X|\textit{``excellent''}])1 + Pr(Y > \mathbb{E}^s[X|\textit{``excellent''}])\mathbb{E}[Y|Y > \mathbb{E}^s[X|\textit{``excellent''}]]\Big]\right)
\end{aligned}$$

which simplifies to

$$\begin{aligned}
\mathbb{E}CS^s(q) &= \frac{1}{2}\left((1-q)\left[\frac{500q^2 - 1200q + 719}{800q^2 - 1920q + 1152}\right] + q\left[\frac{125q^2 + 50q + 4}{200q^2 + 80q + 8}\right]\right) \\
&+ \frac{1}{2}\left(\frac{53}{160} + \frac{1}{5}\left[\frac{600q^2 - 1450q + 875}{800q^2 - 1920q + 1152}\right] + \frac{1}{5}\left[\frac{175q^2 + 80q + 8}{200q^2 + 80q + 8}\right]\right)
\end{aligned}$$

For the expected surplus of a naive consumer we simply replace $\mathbb{E}^s[X|``m"]$ by m and simplify to $\mathbb{E}CS^n(q) = \frac{41-3q}{64}$. Ex-ante consumer surplus is then $\mathbb{E}CS = \frac{1}{2}\mathbb{E}CS^n + \frac{1}{2}\mathbb{E}CS^s$, where ν is the probability that the consumer is naive. The *potential information* *benefit* (PIB) is calculated as follows: A consumer without access to any information about the product would buy it if her outside option is below the expected utility the product yields, which is $\frac{1}{2}$, and otherwise not buy and enjoy her outside option. The expected surplus without access to information is thus given by

$$CS_{min} = Pr(Y \leq \frac{1}{2}) * \mathbb{E}[X] + Pr(Y > \frac{1}{2}) * \mathbb{E}[Y|Y > \frac{1}{2}] = \frac{1}{2} * \frac{1}{2} + \frac{1}{2} * \frac{3}{4} = \frac{5}{8}.$$

A perfectly informed consumer would buy the product if and only if her outside option is below the good's quality. The expected surplus with access to perfect information is thus given by

$$\begin{aligned} CS_{max} = {} & Pr(X = \textit{“very bad”})\,[Pr(Y \leq 0) * u(\textit{“very bad”}) + Pr(Y > 0) * \mathbb{E}[Y|Y > 0]] \\ & + Pr(X = \textit{“bad”}) \left[Pr(Y \leq \frac{1}{4}) * u(\textit{“bad”}) + Pr(Y > \frac{1}{4}) * \mathbb{E}[Y|Y > \frac{1}{4}]\right] \\ & + Pr(X = \textit{“average”}) \left[Pr(Y \leq \frac{1}{2}) * u(\textit{“average”}) + Pr(Y > \frac{1}{2}) * \mathbb{E}[Y|Y > \frac{1}{2}]\right] \\ & + Pr(X = \textit{“good”}) \left[Pr(Y \leq \frac{3}{4}) * u(\textit{“good”}) + Pr(Y > \frac{3}{4}) * \mathbb{E}[Y|Y > \frac{3}{4}]\right] \\ & + Pr(X = \textit{“excellent”})\,[Pr(Y \leq 1) * u(\textit{“excellent”}) + Pr(Y > 1) * \mathbb{E}[Y|Y > 1]] \\ & = \frac{1}{5}\left[0 * 0 + 1 * \frac{1}{2}\right] + \frac{1}{5}\left[\frac{1}{4} * \frac{1}{4} + \frac{3}{4} * \frac{5}{8}\right] + \frac{1}{5}\left[\frac{1}{2} * \frac{1}{2} + \frac{1}{2} * \frac{3}{4}\right] \\ & + \frac{1}{5}\left[\frac{3}{4} * \frac{3}{4} + \frac{1}{4} * \frac{7}{8}\right] + \frac{1}{5}\,[1 * 1 + 0 * 1] = \frac{11}{16}. \end{aligned}$$

The PIB is then given by $CS_{max} - CS_{min} = \frac{1}{16}$. The indirect effect on a sophisticated consumer relative to the PIB is $\frac{\mathbb{E}CS^s(1) - \mathbb{E}CS^s(\frac{\sqrt{43}-2}{5})}{PIB} \approx -0.016$ and that on naives is $\frac{\mathbb{E}CS^n(1) - \mathbb{E}CS^n(\frac{\sqrt{43}-2}{5})}{PIB} \approx 0.065$. The overall effect relative to the PIB is given by $\frac{\mathbb{E}CS^n(1) - \mathbb{E}CS^n(\frac{\sqrt{43}-2}{5})}{PIB} \approx 0.214$.

In order to prove Proposition 1 we will utilize Lemmas $1 - 3$ which we will formulate and prove below.

Lemma 1. *In equilibrium a fake reviewer induces the same posterior with all of the messages he sends:* $\mathbb{E}_T[\mathbb{E}^T[X|m]] = c \ \forall m \in supp(f^F)$. *No higher posterior is induced in equilibrium.*

Proof. Expected purchasing probability is equal to the probability that the consumer's outside option is below the induced posterior expectation. This in turn is equal to the posterior.

$$\begin{aligned}\mathbb{E}_T[Pr(Y \leq \mathbb{E}^T[X|m])] =&(1-\nu)Pr(Y \leq \mathbb{E}^s[X|m]) + \nu Pr(Y \leq \mathbb{E}^n[X|m]) \\ =&(1-\nu)\mathbb{E}^s[X|m] + \nu m\end{aligned}$$

The first equality is due to the linearity of the expectations operator and the second is because Y is distributed uniformly. Hence, a fake reviewer is maximising the induced posterior. The second part of the lemma is straightforward: If some message induced a higher posterior, a fake sender would find it beneficial to send that message instead. □

Lemma 2. *In equilibrium, a fake reviewer mixes whenever there are sophisticated consumers, i.e. whenever $\nu < 1$. His equilibrium strategy is characterised by an atomless distribution f^F.*

Proof. Let us suppose that the fake reviewer played a pure strategy. It is obvious that if he put all probability mass on one message it would be best to put it on the highest one. This indeed is optimal when $\nu = 1$, as the consumer will be naive for sure and buy the good with a probability of 1. If $\nu < 1$, there is positive probability that the consumer is sophisticated. Because real reviewers report truthfully, they send no message with positive probability. A sophisticate will therefore conclude with certainty that $m = 1$ is a fake review and discount it accordingly, while a naive will still take it at face value. Therefore the review will induce a posterior of $(1-\nu)\mathbb{E}[X]+\nu$, while every other message is sent only by real reviewers and therefore induces the correct expectation in both consumer types. At this point it is optimal for a fake reviewer to send some message $m = 1-\epsilon$ because for a small enough ϵ, $(1-\nu)\mathbb{E}[X]+\nu < 1-\epsilon \;\Leftrightarrow\; (1-\nu)(1-\mathbb{E}[X]) > \epsilon$. In fact, a similar argument can be made for any atom in the fake reviewer's message distribution. Because there can be at most countably many atoms in a distribution, for every atom m_i^a there is a message in an arbitrarily small interval $[m_i^a, m_i^a + \epsilon]$ that is not an atom. Because a sophisticate assigns strictly positive probability to it being a real review, it induces a higher posterior than the atom. Therefore, for every atom there is a slightly higher message that constitutes a profitable deviation for the fake reviewer. □

Lemma 3. *Let $\nu < 1$. Then, in equilibrium, a fake reviewer mixes over the interval $[c, 1]$, where c is the posterior that he induces in equilibrium. Furthermore,*

$$c \geq \underline{c} = (1-\nu)\mathbb{E}[X] + \nu > \mathbb{E}[X].$$

Proof. Because the fake reviewer's equilibrium strategy is an atomless distribution f^F, the posterior expectation of a sophisticated consumer in (2) can be written as

$$\mathbb{E}^s[X|m] = \frac{\beta f^F(m)}{1-\beta+\beta f^F(m)}\mathbb{E}[X] + \frac{1-\beta}{1-\beta+\beta f^F(m)}m,$$

which is a convex combination of $\mathbb{E}[X]$ and m. The posterior is therefore also a convex combination of $\mathbb{E}[X]$ and m. Two implications follow from this. Thus, for any review $m < \mathbb{E}[X]$ we have that $\mathbb{E}_T[\mathbb{E}^T[X|m]] < \mathbb{E}[X]$, while for any review $m \geq \mathbb{E}[X]$ we have $\mathbb{E}_T[\mathbb{E}^T[X|m]] \geq \mathbb{E}[X]$. A fake reviewers can therefore not write reviews below $\mathbb{E}[X]$ in equilibrium. This means that $supp(f^b) \subseteq [\mathbb{E}[X], 1]$. Denote by c the constant posterior that fake reviews induce in equilibrium and notice that reviews above $\mathbb{E}[X]$ induce a posterior (weakly) below their face value. This implies $c \leq inf(supp(f^b)) = \underline{m}$. But c cannot be strictly smaller than $\underline{m}$ because otherwise deviating to $\underline{m} - \epsilon$ would be profitable. Hence, $c = inf(supp(f^b))$. Now, to show that the support is in fact the interval $[c, 1]$ and does not have any gaps, let us suppose that there was a message m' with $f^F(m') > 0$ and some $m'' > m'$ with $f^F(m'') = 0$. Because a fake reviewer never sends m'', it induces posterior $\mathbb{E}_T[\mathbb{E}^T[X|m'']] = m''$. But since $c \leq m' < m''$, this is above the constant posterior induced by the fake reviewer, which is ruled out by Lemma 2. Therefore $f^F(m) > 0$ for all $m > c$.

To show the last part of the proposition, note that the lowest posterior that a review $m = 1$ can induce is $\underline{c} = (1-\nu)\mathbb{E}[X] + \nu$, namely if sophisticates are maximally sceptical and discount such a review completely. Thus, a fake reviewer must induce at least a posterior of $\underline{c}$ in equilibrium as otherwise deviating to $m = 1$ would give him a higher payoff. □

Proof of Proposition 1: For $\nu = 1$, fake reviewers always send $m = 1$, As this review induces the highest possible posterior $\mathbb{E}_T[\mathbb{E}^T[X|1]] = 1$, while all others induce a posterior below that, this is the unique equilibrium in that case.

For $\nu < 1$ we can use Lemmas 1 to 3 to find a mixed strategy equilibrium of the game. Given that the fake reviewer mixes according to some density function $f^F(m)$ (Lemma 2) while the real one tells the truth, i.e. he sends each message with unit density, the sophisticated consumer's posterior expectation (2) simplifies to:

$$\mathbb{E}^s[X|m] = \frac{\beta f^F(m)}{1-\beta+\beta f^F(m)}\mathbb{E}[X] + \frac{(1-\beta)}{1-\beta+\beta f^F(m)}m \tag{10}$$

By Lemma 1 we have

$$(1-\nu)\mathbb{E}^s[X|m] + \nu m = c \tag{11}$$

and hence

$$c = (1-\nu)\left(\frac{(1-\beta)}{1-\beta+\beta f^F(m)}m + \frac{\beta f^F(m)}{1-\beta+\beta f^F(m)}\mathbb{E}[X]\right) + \nu m \tag{12}$$

Solving for $f^F(m)$ we get

$$f^F(m) = \frac{1-\beta}{\beta}\frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m} \tag{13}$$

We know from Lemma 3 that $supp(f^F) = [c, 1]$. Furthermore, $f^F(m)$ must integrate to

1. Therefore the following equation pins down the equilibrium value of c:

$$\int_c^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}dm = \frac{\beta}{1-\beta} \tag{14}$$

What we will establish in this proof is that (14) always has a unique solution. For $\beta \in (0,1)$, the RHS of (14) can get arbitrarily large or small. We will show that the same is true for the LHS and that, additionally, it is strictly monotone in c.

We start by showing that

$$\int_c^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}dm \tag{15}$$

can get arbitrarily big. Recall from Lemma 3 the lower bound $\underline{c} = (1-\nu)\mathbb{E}[X]+\nu$. As $c \to \underline{c}$, we have (15) tending to

$$\int_{\underline{c}}^1 \frac{m-(1-\nu)\mathbb{E}[X]-\nu}{(1-\nu)\mathbb{E}[X]+\nu-(1-\nu)\mathbb{E}[X]-\nu m}dm = \int_{\underline{c}}^1 \frac{m-\underline{c}}{\nu(1-m)}dm = \frac{1}{\nu}\int_{\underline{c}}^1 \frac{m-\underline{c}}{1-m}dm$$

Let $h(m) = \frac{m-\underline{c}}{1-m}$. We have to show that $\int_{\underline{c}}^1 h(m)dm$ tends to infinity. Let $\hat{c} \in (c,1)$ and define $g(m) = \frac{\hat{c}-\underline{c}}{1-m}$.

$$g(m)\begin{cases} > h(m) & for\ \ m < \hat{c} \\ < h(m) & for\ \ m > \hat{c} \end{cases}$$

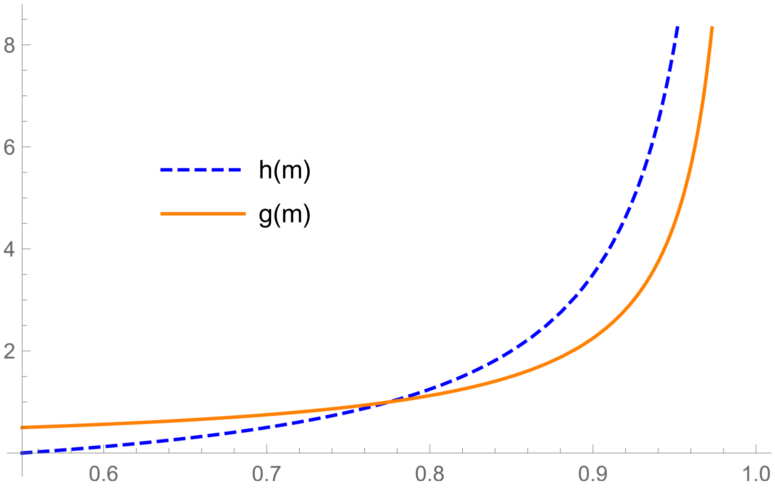

Figure 9: $h(m)$ in blue vs. $g(m)$ in brown

$\int_{\underline{c}}^1 g(m)dm = \infty$ for all $\hat{c} \in (c,1)$ and because $h(m) < g(m)$ only on an interval that contributes a finite part of that integral and $h(m) > g(m)$ on the remaining interval, we must have $\int_{\underline{c}}^{1} h(m)dm = \infty$.

Next, we show that (15) can get arbitrarily small. We can rewrite (15), restricting the range of integration using the indicator function, as

$$\int_0^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}\mathbb{1}_{\{c\leq m\leq 1\}}dm, \tag{16}$$

where the indicator function $\mathbb{1}$ equals 1 whenever the condition inside the braces is satisfied and 0 otherwise. In the limit, as $c \to 1$, we have

$$\lim_{c\to 1}\int_0^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}\mathbb{1}_{\{c\leq m\leq 1\}}dm = \int_0^1 \frac{m-1}{1-(1-\nu)\mathbb{E}[X]-\nu m}\mathbb{1}_{\{m=1\}}.$$

$\frac{m-1}{1-(1-\nu)\mathbb{E}[X]-\nu m}$ equals 0 for $m = 1$ and is finite for $m \in [0, 1)$ (because the denominator is bounded away from 0). We have the integral of a function that is 0 on the entire range of integration and thus

$$\lim_{c\to 1}\int_c^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}dm = 0$$

By the intermediate value theorem, (14) must have a solution for any $\beta \in (0, 1)$.

Finally, we show that (15) is monotonically decreasing in c. Taking the derivative of the left-hand side with respect to c, we have by Leibniz' rule:

$$\frac{d}{dc}\int_c^1 \frac{(m-c)}{c-(1-\nu)\mathbb{E}[X]-\nu m}dm = \int_c^1 \frac{d}{dc}\left(\frac{(m-c)}{c-(1-\nu)\mathbb{E}[X]-\nu m}\right)dm$$

$$= \int_c^1 -\frac{[\overbrace{(c-(1-\nu)\mathbb{E}[X]-\nu m)}^{\geq 0}+\overbrace{(m-c)}^{\geq 0}]}{\underbrace{(c-(1-\nu)\mathbb{E}[X]-\nu m)^2}_{\geq 0}}dm$$

an integral of a function that is negative on the range of integration. Because (15) is monotonically decreasing in c, the solution to (14) is unique. □

Proof of Proposition 2: To show the first part, note that $\int_c^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}dm$ is increasing in ν and decreasing in c. In equilibrium, as (14) needs to hold, these two effects have to cancel out.

For the second part, consider two equilibrium reporting strategies $f^{F,\nu_1}(m)$ and $f^{F,\nu_2}(m)$, with $\nu_1 < \nu_2$. We need to show that $f^{F,\nu_1}(m)$ and $f^{F,\nu_2}(m)$ intersect exactly once in $(c_2, 1)$. First note that $f^F(m) = \frac{1-\beta}{\beta}\frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m}$ is continuous in $]-\infty, 1]$ for $\nu \in [0, 1)$. In particular, $c-(1-\nu)\mathbb{E}[X]-\nu m$ is finite $\forall m \in [\underline{m}, 1]$, $\forall \underline{m} > -\infty$, $\forall \nu \in [0, 1)$. From the first part of the proof we know that $c_1 = c(\nu_1) < c(\nu_2) = c_2$, and because $f^F(m)$ is strictly increasing, $f^{F,\nu_1}(c_2) > f^{F,\nu_2}(c_2)$.

$$\int_{c_2}^{1} f^{F,\nu_1}(m)dm < \int_{c_2}^{1} f^{F,\nu_2}(m)dm = 1$$

, so $f^{F,\nu_1}(m)$ and $f^{F,\nu_2}(m)$ must intersect *at least* once in $(c_2, 1)$. $f^{F,\nu_1}(m) = f^{F,\nu_2}(m)$ yields a quadratic equation in m on $]-\infty, 1]$, so they intersect *at most* twice. $\lim\limits_{m\to-\infty} f^F(m)$, by L'Hospital's rule, is equal to $-\frac{1}{\nu}$ and, hence, $\lim\limits_{m\to-\infty} f^{F,\nu_2}(m) > \lim\limits_{m\to-\infty} f^{F,\nu_1}(m)$. Therefore, $f^{F,\nu_1}(m)$ and $f^{F,\nu_2}(m)$ intersect somewhere in $]-\infty, c_2)$ and thus exactly once in $(c_2, 1)$.
We then have

$$\begin{aligned} f^{F,\nu_1}(m) &> f^{F,\nu_2}(m) \qquad for \quad m < \tilde{c} \\ f^{F,\nu_1}(m) &< f^{F,\nu_2}(m) \qquad for \quad m > \tilde{c} \end{aligned}$$

and therefore

$$\begin{aligned} Pr(M_1 \leq m) \leq Pr(M_2 \leq m) &\Leftrightarrow \int_{c_1}^{\tilde{c}} f^{F,\nu_1}(m)dm \leq \int_{c_1}^{\tilde{c}} f^{F,\nu_2}(m)dm \qquad for \quad m < \tilde{c} \\ Pr(M_1 \geq m) \geq Pr(M_2 \geq m) &\Leftrightarrow \int_{\tilde{c}}^{1} f^{F,\nu_1}(m)dm \geq \int_{\tilde{c}}^{c_1} f^{F,\nu_2}(m)dm \qquad for \quad m > \tilde{c} \end{aligned}$$

□

Proof of Proposition 3: In order to prove Proposition 3, we need to formalise what happens in the limiting case as $\nu \to 1$. Because $c \to 1$ as $\nu \to 1$, $f^F(m)$ converges to a mass point at 1. This is consistent with the fact that when the consumer is naive with probability 1, it is optimal for a fake reviewer to always send the highest message. Formally, $f^F(m)$ converges to a *Dirac delta function* $\delta(m)$ with the properties of being greater than 0 only at $m = 1$ and $\int_{-\infty}^{\infty} \delta(m)dm = 1$. In this sense the Dirac delta function inherits the properties of the sequence of (distribution) functions it is defined as the limit of. Furthermore, $\int_{-\infty}^{\infty} g(m)\delta(m)dm = g(1)$ for all continuous compactly supported functions g. We now turn to the proof.
part (i): receiver surplus of a naive consumer is given by

$$CS^n = \beta \int_{c}^{1} \left([1 - F_Y(m)] \; \mathbb{E}[Y|Y > m] + F_Y(m) \; \mathbb{E}[X] \right) f^F(m)dm$$

$$+ (1 - \beta) \int_{0}^{1} \left([1 - F_Y(m)] \; \mathbb{E}[Y|Y > m] + F_Y(m) \; m \right) dm$$

, which reduces to the following given that Y is uniformly distributed on $[0,1]$:

$$CS^n = \int_c^1 \left[\frac{1-m^2+m}{2}\right] f^b(m)dm + \int_0^1 \left[\frac{1+m}{2}\right] f^u(m)dm.$$

$f^u(m)=1$, so the second integral is independent of ν. By Leibniz' Rule we have

$$\frac{dCS^n}{d\nu} = \int_c^1 \left[\frac{1-m^2+m}{2}\right] \frac{df^F(m)}{d\nu}dm \tag{17}$$

From Proposition 2 we know that $F^{F,\nu_2}(m)$ FSOD $F^{F,\nu_1}(m)$ for $\nu_2 < \nu_1$, i.e. $f^{F,\nu_2}(m)$ and $f^{F,\nu_1}(m)$ intersect once at some $\tilde{c} \in (c_2, 1)$. This means that $\frac{df^F(m)}{d\nu} < 0$ for $m \in (c, \tilde{c})$ and $\frac{df^F(m)}{d\nu} > 0$ for $m \in (\tilde{c}, 1)$. With $\frac{df^F(\tilde{c})}{d\nu} = 0$, we can rewrite (17) as follows

$$\frac{dCS^n}{d\nu} = \int_c^{\tilde{c}} \left[\frac{1-m^2+m}{2}\right] \frac{df^F(m)}{d\nu}dm + \int_{\tilde{c}}^1 \left[\frac{1-m^2+m}{2}\right] \frac{df^F(m)}{d\nu}dm \tag{18}$$

Define $K(m) = \frac{1-m^2+m}{2}$ and note that it is a continuous, strictly decreasing function for $m \in (\frac{1}{2}, 1]$ and that $c > \mathbb{E}[X] = \frac{1}{2}$, by Lemma 2. Then,

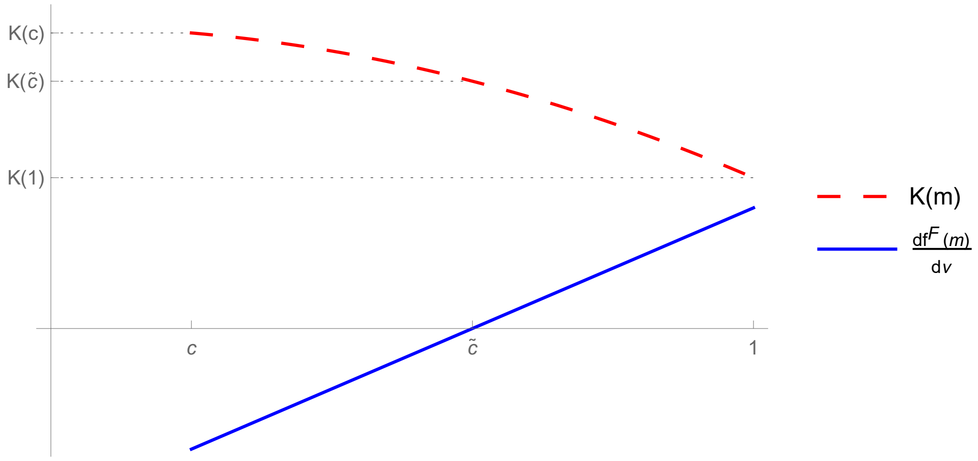

Figure 10: Illustration of proof of Proposition 3(i)

$$K(c)\int_c^{\tilde{c}} \frac{df^F(m)}{d\nu}dm > \int_c^{\tilde{c}} \left[\frac{1-m^2+m}{2}\right] \frac{df^F(m)}{d\nu}dm > K(\tilde{c})\int_c^{\tilde{c}} \frac{df^F(m)}{d\nu}dm$$

and thus $\exists \underline{c} \in (c, \tilde{c})$, such that

$$K(\underline{c})\int_c^{\tilde{c}} \frac{df^F(m)}{d\nu}dm = \int_c^{\tilde{c}} \left[\frac{1-m^2+m}{2}\right] \frac{df^F(m)}{d\nu}dm \tag{19}$$

Similarly,

$$K(\tilde{c})\int_{\tilde{c}}^{1}\frac{df^F(m)}{d\nu}dm > \int_{\tilde{c}}^{1}\left[\frac{1-m^2+m}{2}\right]\frac{df^F(m)}{d\nu}dm > K(1)\int_{\tilde{c}}^{1}\frac{df^F(m)}{d\nu}dm$$

and thus $\exists\overline{c}\in(\tilde{c},1)$, such that

$$K(\overline{c})\int_{\tilde{c}}^{1}\frac{df^F(m)}{d\nu}dm = \int_{\tilde{c}}^{1}\left[\frac{1-m^2+m}{2}\right]\frac{df^F(m)}{d\nu}dm \tag{20}$$

Now $\overline{c} > \underline{c}$ and thus $K(\overline{c}) < K(\underline{c})$. Furthermore,

$$\int_{\tilde{c}}^{1}\frac{df^F(m)}{d\nu}dm = -\int_{c}^{\tilde{c}}\frac{df^F(m)}{d\nu}dm$$

and thus

$$K(\overline{c})\int_{\tilde{c}}^{1}\frac{df^F(m)}{d\nu}dm < -\,K(\underline{c})\int_{c}^{\tilde{c}}\frac{df^F(m)}{d\nu}dm$$

$$K(\underline{c})\int_{c}^{\tilde{c}}\frac{df^F(m)}{d\nu}dm + K(\overline{c})\int_{\tilde{c}}^{1}\frac{df^F(m)}{d\nu}dm < 0$$

Substituting in (19) and (20), we further get

$$\int_{c}^{\tilde{c}}\left[\frac{1-m^2+m}{2}\right]\frac{df^F(m)}{d\nu}dm + \int_{\tilde{c}}^{1}\left[\frac{1-m^2+m}{2}\right]\frac{df^F(m)}{d\nu}dm < 0$$

$$\frac{dCS^n}{d\nu} < 0$$

Note that $c \to 1$ as $\nu \to 1$ and hence in this limiting case we have

$$\lim_{\nu\to 1}\frac{dCS^n}{d\nu} = \lim_{\nu\to 1}\int_{c}^{1}\left[\frac{1-m^2+m}{2}\right]\frac{df^F(m)}{d\nu}dm = \int_{c}^{1}\lim_{\nu\to 1}\left(\left[\frac{1-m^2+m}{2}\right]\frac{df^F(m)}{d\nu}\right)dm$$

$$= \int_{c}^{1}\frac{1}{2}\lim_{\nu\to 1}\frac{df^F(m)}{d\nu}dm = \frac{1}{2}\int_{c}^{1}\lim_{\nu\to 1}\frac{df^F(m)}{d\nu}dm = \frac{1}{2}\lim_{\nu\to 1}\int_{c}^{1}\frac{df^F(m)}{d\nu}dm = 0$$

□

part (ii): A sophisticated consumers ex ante expected surplus is given by

$$CS^s = \beta \int_c^1 \left([1 - F_Y(\mathbb{E}^s[X|m])] \; \mathbb{E}[Y|Y > \mathbb{E}^s[X|m]] + F_Y(\mathbb{E}^s[X|m]) \; \mathbb{E}[X] \right) f^F(m) dm$$

$$+(1-\beta) \int_0^1 \left([1 - F_Y(\mathbb{E}^s[X|m])] \; \mathbb{E}[Y|Y > \mathbb{E}^s[X|m]] + F_Y(\mathbb{E}^s[X|m]) \; m \right) dm,$$

which reduces to the following given that Y is uniformly distributed on $[0,1]$:

$$\begin{aligned} CS^s = \beta \int_c^1 & \frac{1 + \mathbb{E}^s[X|m] - \mathbb{E}^s[X|m]^2}{2} \; f^F(m) \; dm \\ +(1-\beta) \int_0^1 & \left(\frac{1 - \mathbb{E}^s[X|m]^2}{2} \; + \; \mathbb{E}^s[X|m] \; m \right) dm \end{aligned}$$

The derivative with respect to ν is

$$\frac{dCS^s}{d\nu} = \beta \int_c^1 \left(\frac{1 - 2\mathbb{E}^s[X|m]}{2} \; \frac{d\mathbb{E}^s[X|m]}{d\nu} \; f^F(m) + \frac{1 + \mathbb{E}^s[X|m] - \mathbb{E}^s[X|m]^2}{2} \; \frac{df^F(m)}{d\nu} \right) dm$$

$$+(1-\beta) \int_c^1 \left(-\mathbb{E}^s[X|m] \; \frac{d\mathbb{E}^s[X|m]}{d\nu} \; + \; \frac{d\mathbb{E}^s[X|m]}{d\nu} \; m \right) dm$$

Because $\mathbb{E}^s[X|m]$ is given by $\frac{(1-\beta)f^u(m)}{(1-\beta)+\beta f^F(m)} m + \frac{\beta f^F(m)}{(1-\beta)+\beta f^F(m)} \mathbb{E}[X]$, we have

$$\frac{d\mathbb{E}^s[X|m]}{d\nu} = \big(\mathbb{E}[X] - m\big) \frac{\beta(1-\beta)}{[1 - \beta + \beta f^F(m)]^2} \frac{df^F(m)}{d\nu} \tag{21}$$

As $\nu \to 0$ we have $\mathbb{E}^s[X|m] = c$. As shown by Jindapon and Oyarzun (2013) in the limit we have $c = \frac{1}{1+\sqrt{\beta}}$ and

$$f^F(m) = \begin{cases} \frac{2}{\beta}(1+\sqrt{\beta})\big[(1+\sqrt{\beta})m - 1\big] & for \;\; m \in [c,1] \\ 0 & else \end{cases}$$

But then we can rewrite the fraction from (21) as

$$\begin{aligned}
\frac{\beta(1-\beta)}{[1-\beta+\beta f^b(m)]^2} &= \frac{\beta(1-\beta)}{\left[1-\beta+2(1+\sqrt{\beta})\left((1+\sqrt{\beta})m-1\right)\right]^2} \\
&= \frac{\beta(1-\beta)}{\left[(1-\sqrt{\beta})(1+\sqrt{\beta})+2m(1+\sqrt{\beta})^2-2(1+\sqrt{\beta})\right]^2} \\
&= \frac{\beta(1-\sqrt{\beta})(1+\sqrt{\beta})}{(1+\sqrt{\beta})^2\left[(1-\sqrt{\beta})+2m(1+\sqrt{\beta})-2\right]^2} \\
&= \frac{\beta(1-\sqrt{\beta})}{(1+\sqrt{\beta})\left[1-\sqrt{\beta}+2m-2+2m\sqrt{\beta}\right]^2} \\
&= \frac{\beta(1-\sqrt{\beta})}{(1+\sqrt{\beta})\left[(2m-1)+(2m-1)\sqrt{\beta}\right]^2} \\
&= \frac{\beta(1-\sqrt{\beta})}{(1+\sqrt{\beta})^3(2m-1)^2}
\end{aligned}$$

The marginal effect of ν on CS^s in the limit as $\nu \to 0$ is then

$$\begin{aligned}
\lim_{\nu\to 0}\frac{dCS^s}{d\nu} = \beta\int_c^1 &\left(\frac{1}{2}-c\right)\left(\frac{1}{2}-m\right)\frac{\beta(1-\beta)}{(1+\sqrt{\beta})^3(2m-1)^2}\frac{2}{\beta}\left(1+\sqrt{\beta}\right)\left[\left(1+\sqrt{\beta}\right)m-1\right]\frac{df^F(m)}{d\nu} \\
&+\frac{1+c-c^2}{2}\frac{df^F(m)}{d\nu}dm \\
&+(1-\beta)\int_c^1 \frac{(1+\sqrt{\beta})m-1}{(1+\sqrt{\beta})}(\frac{1}{2}-m)\frac{\beta(1-\sqrt{\beta})}{(1+\sqrt{\beta})^3(2m-1)^2}\frac{df^F(m)}{d\nu}dm
\end{aligned}$$

Simplifying and substituting with $c=\frac{1}{1+\sqrt{\beta}}$ we obtain

$$\begin{aligned}
\lim_{\nu\to 0}\frac{dCS^s}{d\nu} = \beta\int_c^1 &\frac{(1-\sqrt{\beta})^2\left[(1+\sqrt{\beta})m-1\right]}{2(1+\sqrt{\beta})^3(2m-1)}\frac{df^F(m)}{d\nu}dm+\underbrace{\beta\int_c^1\frac{1+c-c^2}{2}\frac{df^F(m)}{d\nu}dm}_{=0} \\
&+(1-\beta)\int_c^1\frac{\beta(1-\sqrt{\beta})\left[(1+\sqrt{\beta})m-1\right]}{-2(1+\sqrt{\beta})^4(2m-1)}\frac{df^F(m)}{d\nu}dm
\end{aligned}$$

The second term is 0 because $\frac{1+c-c^2}{2}$ is independent of m and can therefore be pulled

in front of the integral, which equals 0 as established in the preliminary section of the proof.
Finally, pulling both factors inside the integrals and decomposing $1-\beta = (1+\sqrt{\beta})(1-\sqrt{\beta})$, we get

$$\lim_{\nu\to 0}\frac{dCS^s}{d\nu} = \int_c^1 \frac{\beta(1-\sqrt{\beta})^2\left[(1+\sqrt{\beta})m-1\right]}{2(1+\sqrt{\beta})^3(2m-1)}\frac{df^F(m)}{d\nu}dm$$

$$+\int_c^1 \frac{\beta(1-\sqrt{\beta})^2\left[(1+\sqrt{\beta})m-1\right]}{-2(1+\sqrt{\beta})^3(2m-1)}\frac{df^F(m)}{d\nu}dm \; = \; 0$$

In the limiting case as ν tends to 1, we have

$$\lim_{\nu\to 1}\frac{dCS^s}{d\nu} = \beta\int_{c\to 1}^1 \lim_{\nu\to 1}\left[(\frac{1}{2}-\mathbb{E}^s[X|m])\left(\mathbb{E}[X]-m\right)\frac{\beta(1-\beta)\,f^F(m)}{[1-\beta+\beta f^F(m)]^2}\frac{df^F(m)}{d\nu}\right]dm$$

$$+\beta\int_{c\to 1}^1 \lim_{\nu\to 1}\left[\frac{1+\mathbb{E}^s[X|m]-\mathbb{E}^s[X|m]^2}{2}\;\frac{df^F(m)}{d\nu}\;\right]dm$$

$$+(1-\beta)\int_{c\to 1}^1 \lim_{\nu\to 1}\left[(m-\mathbb{E}^s[X|m])\left(\mathbb{E}[X]-m\right)\frac{\beta(1-\beta)}{[1-\beta+\beta f^F(m)]^2}\frac{df^F(m)}{d\nu}\;\right]dm$$

The second term tends to 0 by the same reasoning as in the last part of part(i). The first and third term we can rewrite as

$$\begin{aligned}\frac{dCS^s}{d\nu} =&\beta\int_c^1 \frac{[1-\nu-2(c-\nu m)]\,[c-m+c'(1-\nu)]}{2(1-\nu)^3}f^F(m)dm\\ &+(1-\beta)\int_c^1 \frac{(m-c)\,[c-m+c'(1-\nu)]}{(1-\nu)^3}dm.\end{aligned}$$

Applying de l'Hospital's Rule, it can be shown that the fractions inside both integrals tend to positive but finite numbers as $\nu\to 1$. Because also $c\to 1$ as $\nu\to 1$, the range of integration becomes arbitrarily small and the last term therefore becomes 0. The first one then becomes the integral of something positive but finite multiplied by the Dirac delta function. As shown in the preliminary section of this proof, this must be finite. Hence, we have that $\frac{dCS^s}{d\nu} < \infty$.
part (iii):

Total consumer surplus is given by

$$CS = \nu\, CS^n + (1-\nu)\, CS^s$$

and hence the marginal effect of increasing the share of naive receivers, ν, is

$$\frac{dCS}{d\nu} = (CS^n - CS^s) + \nu \frac{dCS^n}{d\nu} + (1-\nu)\frac{dCS^s}{d\nu}$$

By a simple game-theoretic argument $CS^n - CS^s \leq 0$, because otherwise a sophisticate could imitate a naive consumer and be better off. From part(i) we have that $\frac{dCS^n}{d\nu} < 0$, but for $\frac{dCS^s}{d\nu}$ we have analytical results only for limiting cases. Therefore, the proof of part(iii) will follow part(ii)'s structure, provide analytical proofs for limiting cases and rely on numerical calculations for intermediate parameter values.

$\lim_{\nu \to 0} \frac{dCS^n}{d\nu}$ is negative but finite and hence $\lim_{\nu \to 0} \nu \frac{dCS^n}{d\nu} = 0$.
$\lim_{\nu \to 0} \frac{dCS^s}{d\nu} = 0$ and hence $\lim_{\nu \to 0}(1-\nu)\frac{dCS^s}{d\nu} = 0$.
Thus, $\lim_{\nu \to 0} \frac{dCS}{d\nu} \leq 0$.

$\lim_{\nu \to 1} \frac{dCS^n}{d\nu} = 0$ and hence $\lim_{\nu \to 1} \nu \frac{dCS^n}{d\nu} = 0$.
$\lim_{\nu \to 1} \frac{dCS^s}{d\nu} = 0$ and hence $\lim_{\nu \to 1}(1-\nu)\frac{dCS^s}{d\nu} = 0$.
Because $\lim_{\nu \to 1}(CS^n - CS^s) = -\frac{\beta}{8}$, we then have $\lim_{\nu \to 1} \frac{dCS}{d\nu} < 0$. □

Proof of Proposition 4: Suppose all reviewers and consumers play according to the equilibrium of the baseline model. If a strategically honest reviewer benefits from deviating he must increase the sophisticated consumer's expected welfare. This is because they cannot increase that of a naive consumer - since she take messages at face value, telling the truth is the optimal strategy to maximise her expected welfare. In what follows we will show that by deviating, strategically honest reviewers can trade off small mistakes of a naive against large gains of a sophisticate and thereby increase expected aggregate consumer welfare.

A deviation can only be beneficial if it is upon observing $x > c$ because honest reviews for $x \leq c$ induce the correct posterior expectation in *both* consumer types. Recall, however, that for $m > c$ the posterior expectation of a sophisticated consumer is below the observed quality and furthermore decreasing in m (see Fig. 2). A deviation can only benefit consumers overall if it benefits sophisticated types, hence we need to consider only cases where $\nu < 1$. When $\nu = 0$, the posterior expectation of sophisticates is constant and deviations cannot be beneficial. Therefore, what is left to show is that there exist profitable deviations for $\nu \in (0, 1)$.

Note that upon seeing a truthful review $m \in (c, 1]$, naive consumers always choose the better of the two options. Were the reviewer to underreport, they would instead make some mistakes in expectation. In particular, whenever their outside option lies between the deviation message $m' < m$ and the honest message (and hence true quality) m, they do not make the purchase although they should. For an outside option $y \in [m', m]$ the size of their mistake is $m - y$, i.e. the utility they would have gotten if they optimally bought the good, minus the outside option that they chose instead. Then, for a deviation to some message $m' \in [c, m]$, a naive consumer's expected mistake is given by

$$\int_{m'}^{m} f_Y(\xi)(m-\xi)d\xi = \int_{m'}^{m}(m-\xi)d\xi = m(m-m') - \frac{m^2 - m'^2}{2} = \left(m\xi - \frac{\xi^2}{2}\right)\Big|_{m'}^{m} = \frac{(m-m')^2}{2}.$$

The sophisticates make mistakes when real reviews are truthful, because they discount them. For messages m in the support of the fake reviewer's strategy $[c, 1]$ their posterior expectation is $\mathbb{E}^s[X|m] = \frac{c-\nu m}{1-\nu} < m$. Thus, whenever their outside option is between $\mathbb{E}^s[X|m]$ and m, they do not buy although it would be optimal. If a real reviewer deviated from telling the truth to some $m' \in [c, m)$, a sophisticated consumer would avoid making a mistake of size $m - y$ whenever her outside option was between $\mathbb{E}^s[X|m]$ and $\mathbb{E}^s[X|m']$. Thus her expected *avoided* mistake is given by

$$\int_{\mathbb{E}^s[X|m]}^{\mathbb{E}^s[X|m']} f_Y(\xi)(m-\xi)d\xi = \int_{\mathbb{E}^s[X|m]}^{\mathbb{E}^s[X|m']}(m-\xi)d\xi = \left(m\xi - \frac{\xi^2}{2}\right)\Big|_{\xi=\mathbb{E}^s[X|m]}^{\xi=\mathbb{E}^s[X|m']}$$

To show that a profitable deviation exists, consider a deviation from $m \in (c, 1]$ to c. It is profitable if the expected avoided mistake outweighs the expected mistake:

$$\nu\left(m\xi - \frac{\xi^2}{2}\right)\Big|_{c}^{m} < (1-\nu)\left(m\xi - \frac{\xi^2}{2}\right)\Big|_{\xi=\mathbb{E}^s[X|m]}^{\xi=\mathbb{E}^s[X|c]}$$

$$\nu\left[\left(m^2 - \frac{m^2}{2}\right) - \left(mc - \frac{c^2}{2}\right)\right] < (1-\nu)\left[\left(m\mathbb{E}^s[X|c] - \frac{\mathbb{E}^s[X|c]^2}{2}\right) - \left(m\mathbb{E}^s[X|m] - \frac{\mathbb{E}^s[X|m]^2}{2}\right)\right]$$

Note that $\mathbb{E}^s[X|m] = \frac{c-\nu m}{1-\nu}$ and thus in particular $\mathbb{E}^s[X|m = c] = c$. Then, we have

$$\nu\left[\left(m^2 - \frac{m^2}{2}\right) - \left(mc - \frac{c^2}{2}\right)\right] < (1-\nu)\left[\left(mc - \frac{c^2}{2}\right) - \left(m\frac{c-\nu m}{1-\nu} - \frac{\left(\frac{c-\nu m}{1-\nu}\right)^2}{2}\right)\right]$$

$$\nu\left[\frac{m^2}{2} - \frac{2mc - c^2}{2}\right] < (1-\nu)\left[\frac{2mc - c^2}{2} - \frac{2m\frac{c-\nu m}{1-\nu} - \left(\frac{c-\nu m}{1-\nu}\right)^2}{2}\right]$$

$$\nu\left[\frac{m^2}{2} - \frac{m^2 - (m-c)^2}{2}\right] < (1-\nu)\left[\frac{m^2 - (m-c)^2}{2} - \frac{m^2 - (m - \frac{c-\nu m}{1-\nu})^2}{2}\right]$$

$$\nu\left[\frac{(m-c)^2}{2}\right] < (1-\nu)\left[\frac{(m - \frac{c-\nu m}{1-\nu})^2}{2} - \frac{(m-c)^2}{2}\right]$$

and finally after noting that $m - \frac{c-\nu m}{1-\nu} = \frac{m-\nu m-c+\nu m}{1-\nu} = \frac{m-c}{1-\nu}$, collecting all terms on the LHS, and some rearranging we have

$$\frac{\nu}{(1-\nu)^2}\frac{(m-c)^2}{2} > 0$$

which holds for all $\nu \in (0,1)$ and all $m \in (c,1]$ as assumed above. A deviation from some honest message $m \in (c,1]$ to c is thus profitable for a strategically honest reviewer. □

Proof of Proposition 5: We begin by looking at best responses starting from strategies as in the equilibrium of the baseline model. As shown in the proof of Proposition 4, strategically honest reviewers benefit from deviating to a lower message whenever they observe a quality above c. In fact, their best response is to send $m = c$ for $x \in [c,1]$. Given this strategy, $m = c$ would induce a posterior above c. Sophisticated consumers' posterior expectation would be equal to $\frac{1+c}{2} > c$, naive consumers' posterior expectation would be equal to c and hence the posterior - a convex combination of the two posterior expectations - would lie above c. This would provide an incentive for fake reviewers to deviate to $m = c$ because it induces the highest posterior. Suppose that fake reviewers sent $m = c$ with some probability δ, while they mixed over $[c,1]$ with the remaining probability. Then, the sophisticates' posterior expectation is given by

$$\mathbb{E}^s[X|c] = \frac{\delta\beta}{\delta\beta + (1-c)(1-\beta)(1-\eta)}\frac{1}{2} + \frac{(1-c)(1-\beta)(1-\eta)}{\delta\beta + (1-c)(1-\beta)(1-\eta)}\frac{1+c}{2}$$

Two things need to be true. First, $\mathbb{E}^s[X|c] = c$ so that strategically honest reviewers do not have an incentive to deviate when $x = c$. For $\mathbb{E}^s[X|c] \neq c$ there would be $c-\epsilon$, for small enough *epsilon*, would be such a deviation. Second, $m = c$ has to induce the same posterior as the other messages that the fake reviewer sends. These two facts imply that $\mathbb{E}^s[X|m] = c$ for all $m \in [c,1]$. The first fact yields

$$\frac{\delta\beta}{\delta\beta + (1-c)(1-\beta)(1-\eta)}\frac{1}{2} + \frac{(1-c)(1-\beta)(1-\eta)}{\delta\beta + (1-c)(1-\beta)(1-\eta)}\frac{1+c}{2} = c \tag{22}$$

which can be solved for δ:

$$\delta = (1-\eta)\frac{1-\beta}{\beta}\frac{(1-c)^2}{2c-1} \tag{23}$$

The RHS is continuous and strictly monotonically decreasing in c for $c \in [\frac{1}{2}, 1]$. Thus, (23) implicitly defines a continuous and strictly monotonically decreasing function $c^A(\delta)$, which attains its maximum at $\delta = 0$. Substituting $\delta = 0$ into (23) and solving for c yields $c^A(0) = 1$. As $c^A(\delta)$ is decreasing, we have $c^A(1) < 1$. Figure 11 illustrates $c^A(\delta)$. The second fact implies

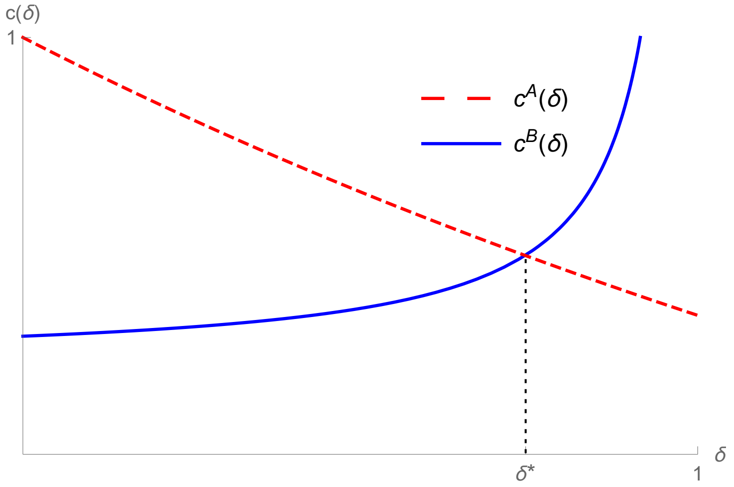

Figure 11: $c^A(\delta)$ and $c^B(\delta)$ intersect exactly once in $[0, 1]$

$$\int_c^1 \eta \frac{1-\beta}{\beta} \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m} dm = 1-\delta \tag{24}$$

It mirrors the equilibrium condition in the baseline model (see eq. (14) in the proof of Proposition 1) but real reviewers write reviews in $[c, 1]$ with density ν instead of 1 and the fake reviewers review density integrates to $1-\delta$ instead of 1. Rearranging gives

$$\int_c^1 \frac{m-c}{c-(1-\nu)\mathbb{E}[X]-\nu m} dm = \frac{1-\delta}{\eta} \frac{\beta}{1-\beta} \tag{25}$$

The RHS of (25) is decreasing in δ and, as shown in the proof of Proposition 1, the LHS is decreasing in c. Thus, (25) implicitly defines a monotonically increasing function $c^B(\delta)$, which attains its maximum at $\delta = 1$. The RHS becomes 0 as $\delta \to 1$ and the LHS becomes 0 as $c \to 1$. Figure 11 illustrates $c^B(\delta)$. We thus have $c^B(1) = 1$ and, because $c^B(\delta)$ is increasing, $c^B(1) < 1$. We now have $c^A(0) > c^B(0)$ and $c^A(1) < c^B(1)$ and thus, by the IVT, $\exists \delta^*$ s.t. $c^A(\delta^*) = c^B(\delta^*)$.

□

Proof of Proposition 6: Suppose $\nu < \frac{1-\sqrt{\beta}}{1-\sqrt{\beta}}$. This is equivalent to $(1-\nu)\mathbb{E}[x] + \nu = \underline{c} < c^{JO} = \frac{1}{1+\sqrt{\beta}}$. Recall that (23) implicitly defines a strictly decreasing function $c^A(\delta)$. Let $\eta \to 0$. Then $c^A(0) = 1$ and $c^A(1) \to= c^{JO}$.

Likewise, recall that (25) implicitly defines a strictly increasing function $c^B(\delta)$. Let $\langle \eta_n \rangle$ be the the corresponding sequence as $\eta \to 0$. Then, for every $\delta < 1$ $\exists N$ s.t. $\forall n > N$ $\frac{1-\delta}{\eta_n}\frac{\beta}{1-\beta} > E$ $\forall E > 0$. Thus, for $\delta < 1$ we have that $\lim_{\eta \to 0} c^B(\delta) = \underline{c}$. Now let $\delta \to 1$ and $\langle \delta_k \rangle$ be the corresponding sequence. $\forall \eta_n$ $\exists K$ s.t. $\forall k > K$ $\frac{1-\delta_k}{\eta_n}\frac{\beta}{1-\beta} < E$ $\forall E > 0$. Thus, for $\delta \to 1$ we have that $\lim_{\eta \to 0} c^B(\delta) \to 1$. All this is illustrated in Figure 12 and implies that in the limit, as $\eta \to 0$, $\delta^* \to 1$ and $c(\delta^*) \to c^{JO}$.

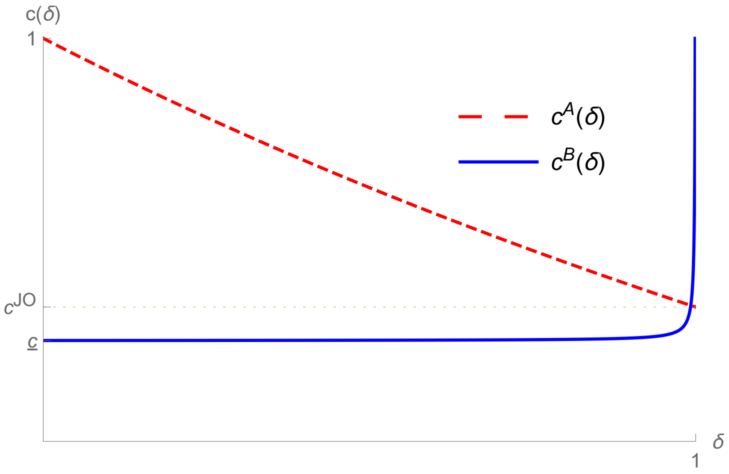

Figure 12: $c^A(\delta)$ and $c^B(\delta)$ as $\eta \to 0$

In the limit there are no naively honest reviewers and $\delta^* \to 1$, which means that reviews above c^{JO} are not sent in equilibrium. Reviews then induce the same posterior expectation in both consumer types, namely $\mathbb{E}^T[X|m] = m$ for $m < c^{JO}$ and $\mathbb{E}^T[X|m] = c^{JO}$ for $m \geq c^{JO}$. A posterior of c^{JO} is thus induced if the reviewer is fake or if he is real and the quality is above c^{JO}. If the reviewer is real and the quality is below c^{JO}, he induces the correct posterior. These are the same posteriors induced in an equilibrium where all reviewers are naively honest and all consumers are sophisticated. □