DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Essays on Information Provision and Consumer Protection“

verfasst von / submitted by

Boris Knapp, BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794 370 140
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Volkswirtschaftslehre
Betreut von / Supervisors:	Univ.-Prof. Dr. Maarten Janssen Univ.-Prof. Philipp Schmidt-Dengler, PhD

Acknowledgements

First and foremost, I want to thank my PhD advisors Maarten Janssen and Philipp Schmidt-Dengler for their mentorship. During the past six years, they have invested great effort in my academic development and struck the right balance between guidance and patience to help me finish this thesis.

I also benefited tremendously from the support of the faculty at the Department of Economics at the University of Vienna. Discussions with Daniel Garcia, Simon Heß, Eeva Mauring, Wieland Müller, Karl Schlag, Juha Tolvanen, Stephanos Vlachos, Geoffrey Castillo, Omar Bamieh, Dijana Zejcirovic, Matan Tsur, and Yan Xu were helpful for guiding my research and for the writing of this thesis.

I want to express my gratitude to Rick Harbaugh for hosting me during a research visit at the Kelley School of Business in Bloomington, Indiana. This was a formative time for me during which I met many interesting people and had the chance to interact with the faculty members of the Department of Business Economics and Public Policy.

My peers at the Vienna Graduate School of Economics were essential during my PhD studies. The many insightful conversations we had, their moral support, and uplifting energy were invaluable for advancing my research and pushing through difficult times. I feel fortunate to have studied alongside such wonderful people, full of humor, empathy, and integrity.

Special thanks go to Dominik Stelzeneder, who co-authored the second chapter of this thesis. The experimental project we carried out was a valuable learning experience and opened up a new field of research for me. Additionally, Richard Nöbauer provided important assistance for the third chapter of this thesis.

Finally, I want to express my endless appreciation for my family. Their unwavering love and support cannot be overstated, and I am forever grateful for their encouragement and guidance throughout this journey. Mom, Dad, Igor, Ana, Petra, Jennifer, I love you!

I would like to acknowledge financial support from the FWF, NOeG, and ÖAW. Their fellowships allowed me to focus on my research and made the completion of this thesis possible.

Contents

Introduction

Information has always been a valuable resource in economics, but it took a while for it to be fully recognized as such. In many ways, Von Hayek's 1945 essay on "The Use of Knowledge in Society" marks the beginning of the field of information in economics. It is one of the earliest thorough examinations of the importance of information and remains a hugely influential paper to this day. However, in 1961, Goerge Stigler wrote: "One should hardly have to tell academics that information is a valuable resource: knowledge *is* power. And yet it occupies a slum dwelling in the town of economics."

One of the main reasons for the initial disregard of its role in economics is that, unlike quantities and prices, information is not easily translated into a mathematical model. However, in the following decades, fundamental progress has been made in this regard, most notably by George Akerlof, Michael Spence, and Joseph Stiglitz, who received the 2001 Nobel Memorial Prize in Economic Science for their groundbreaking contributions.

Today, information has moved out of its former "slum dwelling" and taken center stage in the field of economics. The abundance of data available makes the distribution of information more critical than ever before. As a result, the study of information in economics is experiencing a revival, with implications for businesses, consumers, and regulators. For instance, the regulation of online platforms and information-focused consumer protection are increasingly becoming areas of focus for policymakers.

Against this backdrop, this dissertation provides an exploration of information provision and consumer protection from a variety of angles. It comprises three distinct chapters, each examining the complex relationship between information and regulation policy in different settings.

The first chapter studies the interplay between naive consumers and fake reviews within a theoretical model. Online reviews are vital for many markets, with consumers relying heavily on them to make purchasing decisions. However, firms have a strong incentive to manipulate reviews to increase sales or hurt competitors, resulting in a rise in fake reviews. Consumers' trust in reviews leads to questions about their protection from fake reviews. Through an examination of awareness policies and their impact on both sophisticated and naive consumers, the chapter establishes that these policies inevitably trade off the welfare of one group against the other but ultimately increase aggregate consumer surplus. The chapter also sheds light on the incentives of real reviewers in such markets and establishes that if they write reviews strategically, they are not always truthful.

The second chapter, a collaborative effort with Dominik Stelzeneder, investigates the effects of informational treatments within an asymmetric information game using an online experiment. Asymmetric information is a common occurrence in consumer markets, where producers possess more knowledge about their products than consumers. Customer reviews can mitigate this information asymmetry, but fake reviews introduce a new dimension of information asymmetry that complicates the process of information extraction for consumers. The chapter highlights the challenges associated with misinformation due to fake reviews, which can significantly impair consumers' decision-making abilities by preventing them from computing accurate posterior probabilities. It emphasizes the importance of combining awareness policies with instruments that can help consumers incorporate information correctly, and the need to take gender differences into account.

The third chapter focuses on the effects of an imperfectly enforced rent control policy, where rent caps can only be enforced ex-post and not all tenants are aware of this. It formulates a model of rent control under imperfect enforcement and shows that such regulatory policies can have the unintended effect of increasing rents in the controlled sector. Using data on rental listings in Vienna, the model's hypotheses are tested. The empirical analysis provides partial support for these hypotheses but further research is necessary to provide a more conclusive understanding of imperfectly enforced rent control policies.

The findings of this dissertation have significant implications for policymakers, as they provide valuable insights into the challenges of regulating online platforms and enforcing rent control policies. By shedding light on the complex relationship between information provision and consumer protection, this dissertation contributes to the ongoing discussion of how to best balance the needs of different stakeholders in a rapidly changing world.

Chapter 1

Fake Reviews and Naive Consumers[1]

Consumers very often rely on online reviews when shopping or booking online. I don't want consumers to be tricked. I want them to be able to interact in a trustworthy environment. I insist on one specific point: online businesses must provide consumers with clear and visible information on the reliability of such reviews.

- Didier Reynders, European Commissioner for Justice

1.1 Introduction

Online reviews are a cornerstone of some of today's biggest platforms and have become an integral part of many markets. There are over 250 million reviews on Amazon alone and similar amounts can be found on Google and Yelp. Consumers rely heavily on these reviews to resolve the inherent information asymmetry between them and sellers. Most consumers read online reviews and this affects their purchasing decisions.[2] Firms, therefore, have strong incentives to manipulate reviews in order to boost their sales or hurt their competitors. This is an issue that legislators and platforms alike are trying to crack down on and yet fake reviews are arguably more prevalent than ever.[3] While this is a problem in and of itself it is likely to be exacerbated if consumers are unaware of fake reviews. Unfortunately, there is ample evidence of consumer naivety which raises

[1] A previous version of this chapter was published as Knapp (2020) in the Vienna Economics Papers Working Paper Series.

[2] In a survey conducted by BrightLocal 98% of the respondents read online reviews at least occasionally and according to Podium 93% have made buying decisions based upon an online review. Chevalier and Mayzlin (2006) provide evidence for this for the book market and Lewis and Zervas (2019) find similar results for the accommodation industry.

[3] 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. See also He, Hollenbeck and Proserpio (2022) who provide detailed evidence on how fake reviews are traded on Facebook.

important questions about consumer protection.[4] How can we expect reviewing behavior to be shaped by the presence of naive consumers? What are the effects of awareness campaigns on consumer surplus? Can we expect the market to deal with this problem effectively or is it necessary for policymakers to take an active role?

I study these issues by proposing a cheap-talk model in which a reviewer can be either real or fake and a consumer can be either sophisticated or naive.[5] A consumer considers buying a product of unknown quality and reads a review in order to make an informed decision. If the consumer is of the naive type she takes the review at face value. This is not necessarily true for the sophisticated type who rationally takes into account the possibility that the review might be fake. To start, I solely focus on the fake reviewer's incentives to write reviews strategically in order to maximise the purchasing probability and therefore the baseline model assumes the real reviewer always writes a truthful review. I show that, in equilibrium, fake reviews blend in with real ones as the fake reviewer balances maintaining credibility vis-a-vis sophisticated consumers and persuading naive ones. This trade-off shifts when consumers become increasingly aware because the fake reviewer attaches greater importance to the aware consumer group. As a result, fake reviews are buried deeper in the review pool, which makes reviews less informative for aware consumers but fake reviews less deceptive for naive ones. Therefore, awareness policies do not only have the direct effect of making some naive consumers sophisticated but they also affect all remaining consumers via the changes in the fake reviewers' strategy. The naive consumers who are made aware are better off, those who remain naive benefit because reviews are less deceptive, and the sophisticated consumers suffer because they are less informative. The positive effects always outweigh the negative ones so a policymaker should always try to raise more awareness.

These findings are important for two reasons. First, despite the overall benefit of awareness policies they trade off the welfare of different consumer groups and are thus non-trivial from the perspective of consumer protection. Second, because the aware consumers are the ones who suffer from such policies it is unlikely that they raise awareness. Likewise, a platform does not only risk damaging its reputation but also forgoes revenues by raising awareness. The latter is true because the expected purchasing probability increases with the share of naive consumers which in turn increases the profits of a platform that collects

[4]The literature on strategic communication has documented excessive trust times and times again (Cai and Wang, 2006; Kawagoe and Takizawa, 2009). In a signaling context Deversi et al. (2021) provide experimental evidence that a majority of receivers are oblivious to the strategic incentives of senders. Surveys also document that consumers trust online reviews (CPC Strategy, 2019; Bright Local, 2020).

[5]I model sophistication and naivety in the spirit of Ottaviani and Squintani (2006). Naivety thus captures the fact that consumers are oblivious and differs from the concept of naïveté in the work of Heidhues and Kőszegi (2010, 2017).

fees. Therefore, the problem of naive consumers is unlikely to be taken care of by the market which means that a policymaker needs to assume an active role in dealing with this issue.

I then allow the real reviewer type to be strategic I show that in order to maximise consumer surplus he does not always write a truthful review. Instead, he understates the product's quality when it is high in order to avoid being perceived as fake by aware consumers. *Underreporting* is optimal because the welfare gain of aware consumers outweighs the loss of naive ones. This is true even when the share of aware consumers is very small. Finally, I show that as more and more real reviewers become strategic and the share of unstrategic ones vanishes, the outcome is equivalent to that of an equilibrium with unstrategic real reviewers where the share of naive consumers vanishes. This result is of particular relevance in markets where the boundary between reviewers and consumers is fluid as is the case for most online platforms. It implies that if awareness policies have the effect of making real reviewers strategic, they are effective irrespective of which side of the market they reach.

While this paper deals mainly with fake reviews its findings also apply to other contexts with similar informational and strategic environments. 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 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 consumer naivety might play a role here as well. Awareness campaigns might therefore be suitable policy instruments in these markets as well. The findings in this paper provide important theoretical underpinnings for this discussion.

While the model analysed in this paper remains agnostic regarding the details of awareness policies, I want to briefly discuss several possible implementations. The quote at the beginning of this section suggests an approach that relies on platforms informing

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

consumers. In it, the European Commissioner for justice insists that businesses provide visible information about the reliability of reviews. Such information could for example be displayed at the top of each page that contains reviews. This approach is likely to be both effective and efficient, displaying the relevant information to the consumers who are affected by fake reviews when they are affected. Awareness policies could also be implemented by means of media campaigns. In 2016, the European Commission launched a two-week information campaign on Facebook to raise attention about consumer rights. Similar campaigns providing information about fake reviews would have the effect of raising consumers' attention to the issue. Lastly, more indirect measures also have the potential to increase consumer awareness. For example, public authorities could support research about fake reviews while also fostering media coverage thereof. While such measures are less immediate than the ones discussed previously, they would lead to more awareness as information eventually spreads among consumers.

The present work is most closely related to Kartik, Ottaviani and Squintani (2007), which is the only other paper that studies welfare effects of awareness policies in a sender-receiver game.[7] The main difference between this paper and Kartik et al. (2007) is that I model the sender's bias along the extensive margin. That is, a sender is either biased, in which case his preferences are completely misaligned with those of the receiver, or he is not, in which case they are perfectly aligned. In contrast to that, Kartik et al. (2007) model bias along the intensive margin such that the sender is always biased but his preferences are partially aligned with those of the receiver. In the context of online reviews, the extensive margin is the more important one since fake reviews arguably do not take consumers' well-being into consideration and because not all reviews are fake. This difference leads to vastly different findings and in turn opposing policy recommendations. While awareness policies in this paper increase overall receiver surplus and particularly help naive receivers, they hurt naive receivers and have no effect on sophisticated ones in Kartik et al. (2007). Their findings hinge on the fact that sophisticated receivers, having perfect knowledge of the sender's motives, can de-bias deflated equilibrium messages. The sender's incentive to inflate messages beyond what sophisticated receivers already de-bias is offset by the fact that naive receivers who follow these same messages take actions that are suboptimal even for the sender. Naive receivers, hence, suffer a lot while sophisticated ones enjoy perfect information. Awareness policies that reduce the share of naive receivers decrease the cost to the sender of inflating language which leads to more inflated language in equilibrium and an even lower surplus for naive receivers. In my model, the receiver is uncertain about the sender's type and, in equilibrium, the biased sender pools in order

[7]Another related paper is Ottaviani and Squintani (2006) which differs from Kartik et al. (2007) only in that it assumes a bounded state space.

to sustain this uncertainty. While the biased sender's equilibrium messages are inflated, they become more moderate when the share of naive receivers shrinks. While there are environments where the modeling choices in Kartik et al. (2007) are more appropriate, I believe that my model better captures the strategic interaction in a market with online reviews. The strikingly different findings in this paper and Kartik et al. (2007) highlight the importance of paying attention to the details of a market as to not follow misguided conclusions.

Several other papers study communication games with heterogeneity on the sender side (Jindapon and Oyarzun, 2013; Glazer, Herrera and Perry, 2021), 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 strategically honest senders. The only exception is Glazer et al. (2021) who consider strategically honest senders. However, in the absence of oblivious consumers they do not have an incentive to deviate from truthtelling.

More broadly, this paper contributes to the literature in strategic communication. The seminal work of Crawford and Sobel (1982) has been extended 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; Chakraborty and Harbaugh, 2010), 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). More recently, Kamenica and Gentzkow (2011) proposed a model of persuasion which assumes that the sender can commit ex-ante to a communication strategy. Lipnowski and Ravid (2018) analyse communication in the cheap talk framework, i.e. without ex-ante commitment, and compare to Kamenica and Gentzkow (2011)'s persuasion framework.

The underreporting result in this paper relates to *political correctness* in Morris (2001), *countersignaling* in Feltovich et al. (2002), and *reversal* in Smirnov and Starkov (2022). While similar in flavor, all four results describe distinct phenomena that arise due to different strategic incentives.

In order to focus on the details of communication, I abstract from the pricing stage. Recent papers have studied the role of pricing in the presence of reviews (Martin and Shelegia, 2021) or soft product information (Janssen and Roy, 2022).

The remainder of the paper is structured as follows: An illustrative example is provided in the following section. In section 3, we formalise the baseline model where real reviews are always truthful and in section 4 we characterise its equilibrium. In section 5, we study the effects of educational policies by analysing the comparative statics of the model. We then look at extensions of the baseline model. In section 6, we allow for fake reviews to be either positive or negative and show that our findings remain unchanged. In section 7, we extend the model to multiple consumers and reviewers, and in section 8, we introduce strategically honest reviewers. Finally, we conclude in section 9.

1.2 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 review is a message from the set of possible messages, $\mathcal{M} = \{1\star, 2\star, 3\star, 4\star, 5\star\}$. There is a commonly understood relationship between messages and quality such that $1\star$ corresponds to the worst and $5\star$ to the best quality level. 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 $\frac{3}{5}$ and suppose that a fake reviewer always sends a $5\star$ review. 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:

$$\begin{aligned} Pr(fake|5\star) &= \frac{Pr(5\star|fake) * Pr(fake)}{Pr(5\star|fake) * Pr(fake) + Pr(5\star|real) * Pr(real)} \\ &= \frac{1 * \frac{1}{2}}{1 * \frac{1}{2} + \frac{1}{5} * \frac{1}{2}} = \frac{5}{6} \end{aligned}$$

Her posterior expectation conditional on observing a $5\star$ review is then given by:

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

Hence, upon observing a $5\star$ review, she will purchase the good with probability $\frac{7}{12}$. The expected purchasing probability that a $5\star$ review induces is then $Pr(purchase|5\star) =$

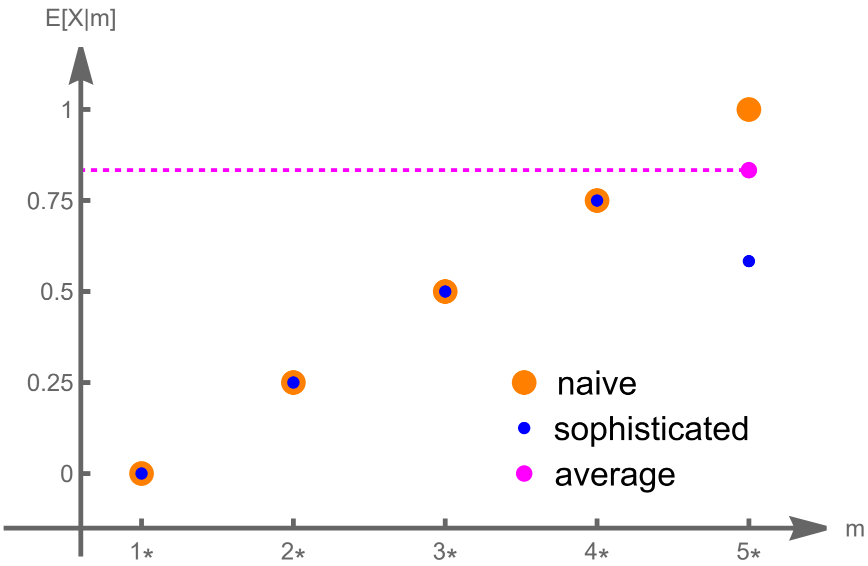
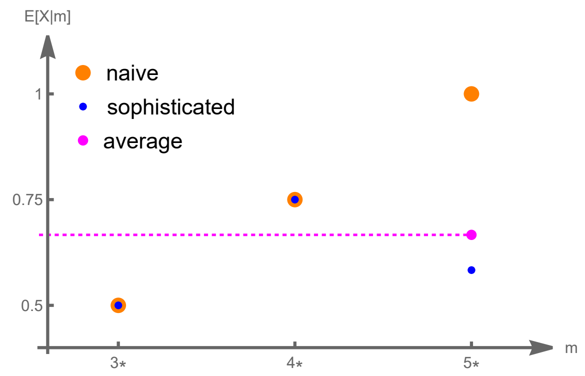


Figure 1.1: Posterior expectations and expected purchasing probability when fake reviewers always send a 5⋆ review

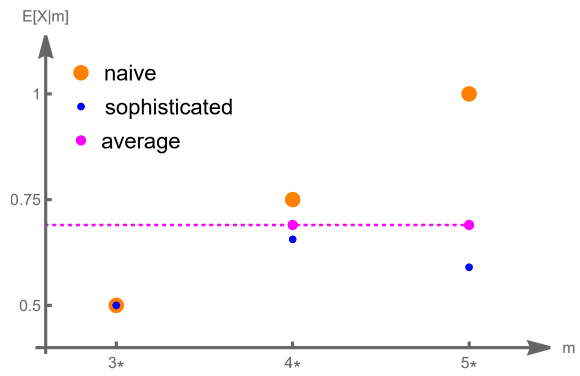
$\frac{1}{2} * 1 + \frac{1}{2} * \frac{7}{12} = \frac{19}{24} > \frac{3}{4}$. Because all other messages induce lower expected purchasing probabilities, it is optimal for the fake reviewer to pursue the pure strategy we assumed in the beginning. This situation is illustrated in Figure 1.1.

Consider now an awareness campaign that reduces the share of naives from $\frac{3}{5}$ to $\frac{1}{5}$. Suppose that, as before, a fake review is always a 5⋆ review. The expected purchasing probability would then decrease to $Pr(purchase|5\star) = \frac{1}{5} * 1 + \frac{4}{5} * \frac{7}{12} = \frac{2}{3}$. This means that deviating to claiming *"good"* instead of *"excellent"* quality is more likely to induce a purchase. As a result, the pure strategy from before is no longer optimal. Instead, a fake reviewer must mix over 4⋆ and 5⋆ reviews in such a way that both induce the same purchasing probability (see Figure 1.2). In our example, the optimal mixed strategy is to send a 4⋆ review 11% of the time and a 5⋆ review the remaining 89% of the time.

This adjustment of the fake reviewer's strategy affects the welfare of both consumer types. For a naive consumer, a 4⋆ review is less deceptive because it induces a belief that is closer to the unconditional expectation. Note that fake reviews are uninformative. Hence, such a belief is more accurate and induces undesirable purchases less often. The *deception effect* is thus positive. For sophisticated consumers, information is lost because a 4⋆ review can no longer be trusted, and because the expected quality conditional on a 4⋆ review is more similar to that conditional on a 5⋆ review. The *separation effect* is thus negative. Consequently, the two consumer groups are affected in opposite ways by the awareness campaign. As we will see in the remainder of this paper, these effects extend to a continuous message space and also arise when fake reviewers try to minimize the purchasing probability.



(a) After the awareness campaign, the purchasing probability falls below $\frac{3}{4}$...



(b) ...which induces the fake reviewer to mix between 4⋆ and 5⋆

Figure 1.2: Effect of an awareness campaign on a fake reviewers equilibrium strategy

1.3 Model

There are two players, a sender, called reviewer, and a receiver, called consumer. Before choosing between a good of unknown quality and her outside option, the consumer reads the reviewer'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 consumer 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 reviewer can be of two types, real or fake, which differ along two dimensions. First, they differ in terms of information. Neither reviewer type knows the realisation of the consumer's outside option, but while the real type observes the good's quality the fake one does not.[8] Second, they differ with regard to their strategic incentives. The real reviewer is a non-strategic player who always writes a truthful review, i.e. he passes on his private information honestly.[9] The fake reviewer is strategic with the objective of 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 consumer'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 consumer's outside option. Note that a fake reviewer cannot convey any information with his review since he does not observe

[8] 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. (2021) to accommodate both cases. The equilibrium strategy that we derive is optimal also for a fake reviewer who observes the quality.

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

the good's quality.

The consumer also can be of two types, sophisticated or naive. A sophisticated consumer 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 consumer on the other hand takes every message at face value. The consumer'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 $(a = 0)$.[10]

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

$$\mathbb{E}^n[X|m] = m. \tag{1.1}$$

Sophisticated consumers take into account that the review might come from a fake reviewer 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 reviewer is fake given that m was sent times the unconditional expected quality (because fake reviews are uninformative), plus the probability that the reviewer is real given m times m (because real reviews are truthful):

$$\mathbb{E}^s[X|m] = Pr(fake|m)\mathbb{E}[X] + (1 - Pr(fake|m))\, m \tag{1.2}$$

The consumer 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 consumer's choice depends on the review m and the realisation y, we can formulate her choice function as

$$a^T(y, m) = \begin{cases} 1 & \mathbb{E}^T[X|m] \geq y \\ 0 & \mathbb{E}^T[X|m] < y \end{cases}, \tag{1.3}$$

where $T \in \{s, n\}$ represents the consumer'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 consumer being naive, $\nu \in (0, 1)$, and of the reviewer 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:

[10] Note that we abstract from price setting behavior in this paper. Thus, our results are of particular relevance in markets where prices are either fixed or can only be adjusted in the long run, while reviews represent a short-term interaction. The interplay between consumer deception and price setting is explored in Janssen and Roy (2022).

1. Nature draws x, y, and a type for the consumer and reviewer.
2. The reviewer observes his type and - if he is real - also the good's quality. He then sends a review, $m \in [0, 1]$, to the consumer.
3. The consumer 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 reviewer maximises the expected purchasing probability, given his (prior) belief about the consumer's type. A sophisticated consumer maximises her payoff taking the reviewer's strategy as given and given her beliefs about his type. A naive consumer maximises her payoff taking every review at face value. Formally, a PBE of the game is a pair of purchasing strategies for the two consumer types and a reporting strategy for the fake reviewer 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 consumer type, which are consistent with Bayes' Rule and such that her strategy is optimal given these beliefs.

1.4 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 consumer type to maximise the expected surplus given her beliefs. According to (c), a biased reviewer maximises the expected purchasing probability. The subscripts Y and T denote that expectations are taken over both the consumer'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. To simplify notation we define $q(m, \beta, \nu) := \mathbb{E}[X|m]$ and use the shorthand notation $q(m)$ throughout. We also denote $\mathbb{E}[X] = \bar{q}$.

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

Proposition 1 *There exists a unique Perfect Bayesian Equilibrium for any share of fake reviewers, β, and any share of naive consumers, ν. The reporting strategy of the fake reviewer is given by $f^F(m) = \frac{1-\beta}{\beta}\frac{m-c}{c-(1-\nu)\bar{q}-\nu m}$ where c is the posterior that fake reviews induce in equilibrium and the unique solution to $\int_c^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m} dm = \frac{\beta}{1-\beta}$.*

Proof. See Appendix 1.B.

To better see what the fake reviewer's equilibrium strategy looks like and understand the intuition behind the equilibrium, it is helpful to look at Figures 1.3 and 1.4. Figure 1.3 depicts the review distributions of the two reviewer types. Because the honest reviewer is truthful and the good's quality is distributed uniformly, so are his reviews (dashed gray). The fake reviewer only sends reviews above a threshold value c (solid red). Because he wants to convince the consumer 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 reviewer writes no single review with positive probability. Therefore, were the fake reviewer to always claim the highest quality in equilibrium, the sophisticated consumer type could infer his type with certainty and ignore the review. Hence, while such a review would be very effective in persuading a naive consumer 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 reviewer 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 reviewer puts more probability mass on reviews the higher they are. The increased scepticism of the sophisticated consumer is compensated for by the naive's belief that it is truthful. Figure 1.4 shows the posterior beliefs of both consumer 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 (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

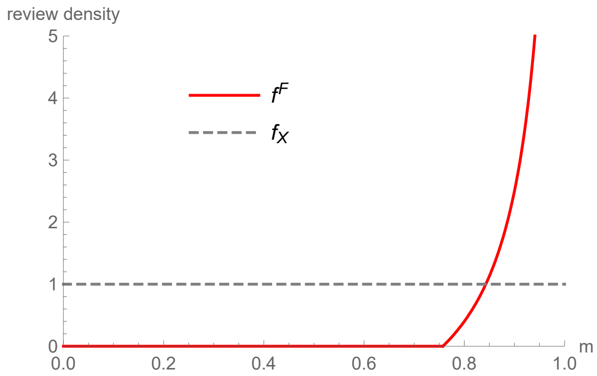


Figure 1.3: Equilibrium message distributions, $\beta = 0.5$, $\nu = 0.5$

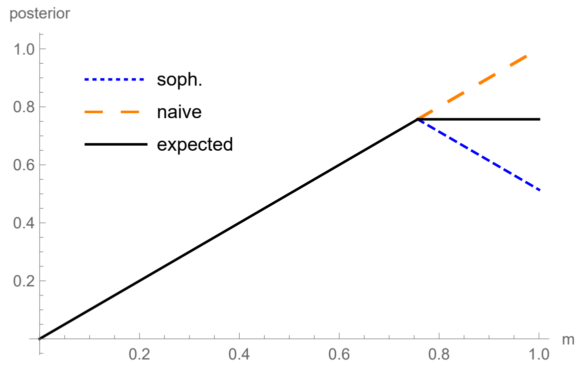


Figure 1.4: Equilibrium posterior beliefs, $\beta = 0.5$, $\nu = 0.5$

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 consumer surplus change as a result of educational policies.

1.5 The Effect of Educational Policies

As outlined in the previous section, a fake reviewer 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 1.3 and 1.4, the higher the density that a fake reviewer puts on a message, the more that message gets discounted by a sophisticated consumer.

Now, if the consumer is more likely to be naive, then the fake reviewer 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 reviewer's strategy in the sense of first order stochastic dominance. Moreover, the expected posterior and thus also the purchasing probability increases as the consumer is more likely to be naive.

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

Proof. See Appendix 1.B.

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 consumer. The increased scepticism is thus outweighed by the fact that it is less likely to occur. We can see in Figure 1.5 how the fake reviewer's distribution shifts as the share of naive consumer increases from $\frac{1}{2}$ to $\nu = \frac{3}{4}$.

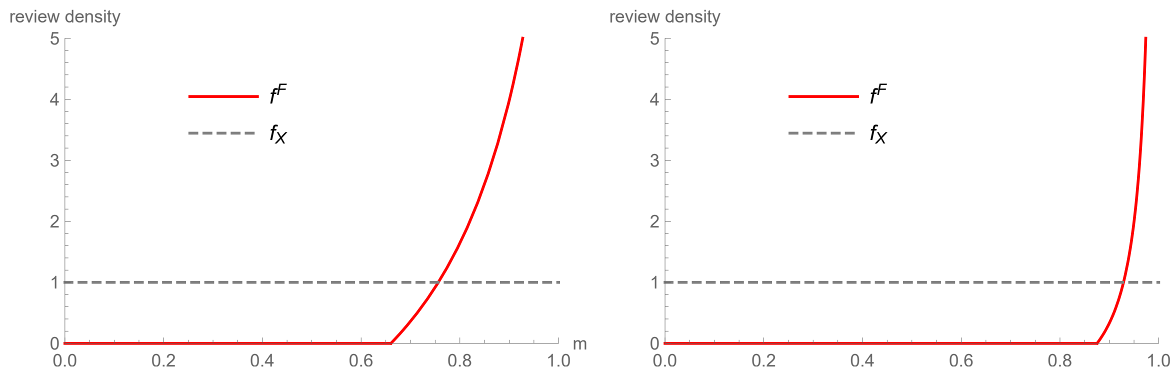


Figure 1.5: equilibrium message distributions for $\nu = 0.25$ (left) and $\nu = 0.75$ (right)

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 reviewers 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.

1.5.1 Consumer Surplus

As pointed out in the introduction, consumer welfare is of particular importance to policymakers. 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:

$$\begin{aligned} CS^n =& \beta \int_c^1 \left(\left[1 - F_Y(m)\right] \mathbb{E}[Y|Y > m] + F_Y(m)\, \bar{q} \right) f^F(m) dm \\ &+ (1-\beta) \int_0^1 \left(\left[1 - F_Y(m)\right] \mathbb{E}[Y|Y > m] + F_Y(m)\, m \right) dm \end{aligned} \tag{1.4}$$

Taking a closer look at (1.4), 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 $\bar{q}$ 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.

$$\begin{aligned} CS^s =&\beta \int_c^1 \big(\left[1 - F_Y(q(m))\right] \mathbb{E}[Y|Y > q(m)] + F_Y(q(m))\, \bar{q} \big)\, f^F(m) dm \\ &+ (1-\beta) \int_0^1 \big(\left[1 - F_Y(q(m))\right] \mathbb{E}[Y|Y > q(m)] + F_Y(q(m))\, m \big)\, dm \end{aligned} \tag{1.5}$$

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

$$CS = \nu\, CS^n + (1-\nu)\, CS^s. \tag{1.6}$$

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:

$$\frac{dCS}{d\nu} = (CS^n - CS^s) + \nu \frac{dCS^n}{d\nu} + (1-\nu)\frac{dCS^s}{d\nu}. \tag{1.7}$$

An *increase* in the 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 (1.7) 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 terms correspond to the *indirect effects* on the naive and sophisticated consumer group, respectively.

The first term is always (weakly) negative. Were a naive consumer to enjoy a higher surplus, a sophisticate could simply imitate them 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 reviewer'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 reviewers 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 reviewer 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 reviewer 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.

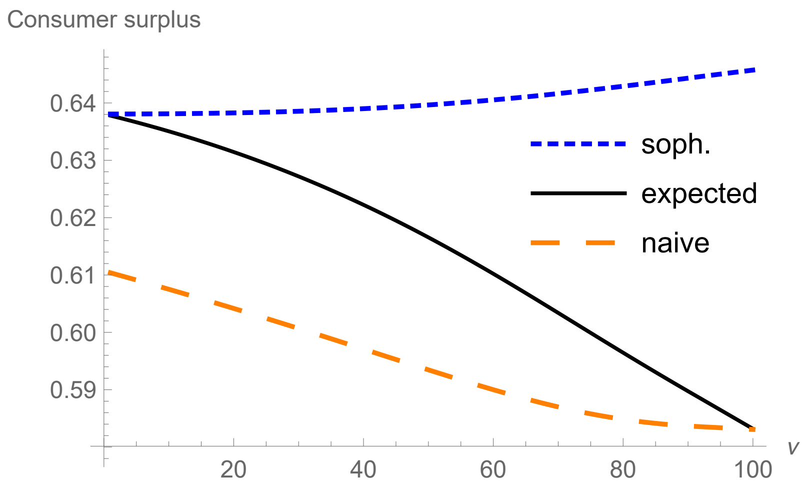


Figure 1.6: naive, sophisticated, and aggregate consumer surplus as a function of the share of naive consumers

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 policymaker should naturally consider the effect on aggregate consumer welfare. Proposition 3 summarizes our findings about educational policies and establishes that they always benefit consumers in aggregate.

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

(i) naive consumers are harmed,

(ii) sophisticated consumers benefit,

(iii) and aggregate consumer surplus decreases.

Proof. See Appendix 1.B.

Figure 1.6 illustrates this. It shows how consumer surplus for naives (top left), sophisticates (top right), and aggregate consumer 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 is 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.

1.6 Negative Fake Reviews

So far we have only considered positive fake reviews, i.e. fake reviews with the objective of increasing sales. In practice, however, fake reviews can be both positive and negative. Mayzlin et al. (2014) not only show that hotels with higher incentives to manipulate reviews have more positive ones but also that hotels whose neighbours have higher incentives for manipulation have more negative reviews. This shows that negative fake reviews by rivals are a problem as well. To account for this, we extent the baseline model to allow for both types of fake reviews. Let β_H be the probability that a review is a positive fake review and β_L the probability that it is a negative fake review. With a probability of $1 - \beta_H - \beta_L$ a review is real and thus truthful. Additionally, we replace equilibrium condition (c) with conditions $(c.1)$ and $(c.2)$.

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

(c.2) $m_L^{F*} \in \arg\min_{m\in[0,1]} \mathbb{E}_{Y,T}[a^{T*}(m, Y)]$
for all $m^* \in supp(f_L^F)$

We can then characterize the fake reviewers' equilibrium strategies by extending Proposition 1 to:

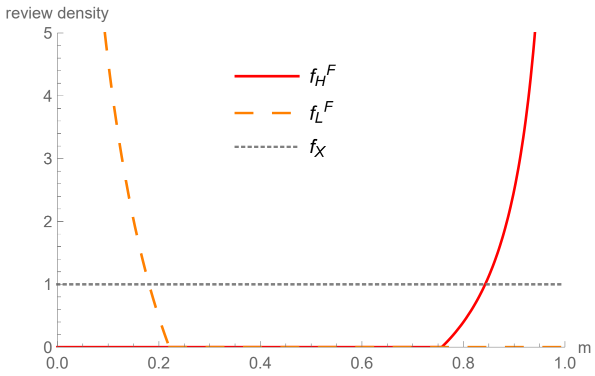


Figure 1.7: Equilibrium message distributions, $\beta_H = 0.5$, $\beta_L = 0.25$, $\nu = 0.5$

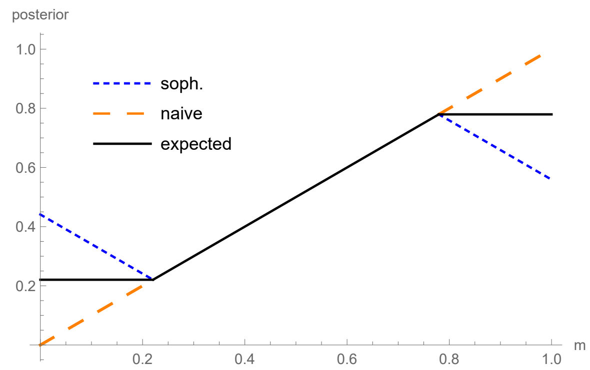


Figure 1.8: Equilibrium posterior beliefs, $\beta_H = 0.5$, $\beta_L = 0.25$, $\nu = 0.5$

Proposition 1b *There exists a Perfect Bayesian Equilibrium for any shares of fake reviewers, β_H and β_L, and any share of naive consumers, ν. The reporting strategy of the fake reviewer is given by $f_i^F(m) = \frac{1-\beta_i}{\beta_i}\frac{m-c_i}{c_i-(1-\nu)\bar{q}-\nu m}$, for $i \in \{L, H\}$. c_H is the posterior that positive fake reviews induce in equilibrium and the unique solution to $\int_{c_H}^1 \frac{m-c_H}{c_H-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta_H}{1-\beta_H}$. c_L is the posterior that negative fake reviews induce in equilibrium and the unique solution to $\int_0^{c_L} \frac{m-c_L}{c_L-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta_L}{1-\beta_L}$.*

Proof. See Appendix 1.C.

One thing to notice is that the positive and negative fake reviewers' strategies depend on one another only mechanically via β. An increase in e.g. β_H, ceteris paribus (in particular, holding β_L constant), increases the ratio of negative fake reviewers to real ones thus affecting the negative type's strategy via $\frac{\beta_L}{1-\beta}$. Beyond that, however, there is no interdependence. In particular, changing β_H while holding the ratio of negative fake reviewers to real ones constant does not affect the negative type's equilibrium strategy despite changing the positive type's equilibrium behavior. In this sense, the two fake types are *strategically independent* from one another. To understand this, recall that a positive fake reviewer never sends messages below the unconditional expectation $\bar{q}$. Similarly, a negative fake never sends messages below $\bar{q}$ and therefore the two fake reviewer types do not interfere with each other. Graphically, this yields a symmetric counterpart on the low quality side of the quality space as shown by Figures 1.7 and 1.8.[11]

Intuitively, our findings about educational policies (Proposition 3) should hold with the inclusion of negative fake reviewers because they affect consumers in a similar way to how positive fake reviewers did in the baseline model. Naive consumers suffer from

[11]The exact mirror image only obtains for $\beta_H = \beta_L$. In general we have a "qualitatively symmetric" counterpart.

deception because negative fake reviewers make them forgo purchases, while sophisticated consumers suffer from a loss of separation as they can no longer fully trust low reviews. To show that this is indeed the case, we prove Proposition $3b$ in Appendix 1.B.[12]

Because the inclusion of negative fake reviewers offers little additional insight but comes at the cost of cluttered notation, we continue our analysis with only positive fake reviews from here onwards.

In this section, we assumed that consumers do not differ in their naivety towards positive and negative fake reviews. We could easily allow for such a distinction by distinguishing the share of consumers who are naive towards positive fakes, say ν_H, from that of consumers who are naive towards negative fakes, say ν_L. This would merely introduce more notation but leave all results unchanged.

1.7 Multiple Reviewers and Multiple Consumers

For the sake of parsimony, we formulated our model as the interaction between one reviewer and one consumer. This is a drastic simplification, as in reality, multiple consumers read multiple reviews before making their purchasing decisions. A recent survey conducted by digital.com revealed that 88% of online shoppers read at least three reviews before a purchase.[13] In this section we will do away with this simplification and allow for multiple consumers and multiple reviewers by replacing the single consumer with a unit mass of consumers and by allowing $k \geq 1$ reviewers to each post a review.

The reason to replace the single consumer with a mass instead of a finite number of consumers is twofold. First, it reflects the fact that the number of consumers typically far exceeds the number of reviews. Second, it is the easiest way to relax the single consumer assumption without complicating the model. In fact, we only need to reinterpret our existing model. Instead of interpreting ν as the probability that the consumer be naive, we interpret it as the share of naive consumers. Furthermore, the expected posterior becomes the average posterior and the purchasing probability becomes the realised sales. This leaves the fake reviewer's incentives and all findings unchanged.

Extending the model to allow for multiple reviewers is a more substantial modification that requires additional analysis. We do not intend to characterize the set of all equilibria in this section but instead prove that an equilibrium exists. We do this by chcharacterising he so-called *negative assortative equilibrium*, which is equivalent to the equilibrium in the baseline model in a sense to be clarified later in this section. For this equilibrium to exist, we need to modify our model beyond allowing for multiple reviewers. First, we need to

[12] We do not state the proposition here as it is identical to Proposition 3.

[13] See https://digital.com/54-of-online-shoppers-read-reviews-before-every-purchase/

assume that fake reviewers have access to some kind of randomisation device that allows them to coordinate their reviews. We let the product quality serve as a coordination device and allow fake reviewers to observe it. Note that this is in contrast to our baseline model where we assumed that only real reviewers observed product quality. Second, we need to specify consumers' out-of-equilibrium beliefs. The reason that this was not necessary for the baseline model is that any deviation was an *on-schedule* deviation, meaning that it could also have been sent in equilibrium. This made any out-of-equilibrium review indistinguishable from equilibrium reviews and the belief updating process fully specified a consumer's beliefs. With multiple reviewers, even though any single review is consistent with equilibrium behavior, there may be combinations of reviews that are not. We call such deviations that result in combinations that are incompatible with equilibrium behavior *off-schedule* deviations and assume that consumers update beliefs pessimistically when confronted with an off-schedule out-of-equilibrium review combination. More specifically, we make the following assumption.

Assumption A1 Let $\mathbf{m} = (m_1, m_2, ..., m_k)$ denote a vector of k reviews. Then a consumer's off-schedule out-of-equilibrium beliefs are given by $q(\mathbf{m}) = \min(m_1, m_2, ..., m_k)$.

These out-of-equilibrium beliefs permit the following equilibrium.

Proposition 4 (Negative Assortative Equilibrium) *Under Assumption A1, there exists a Negative Assortative Equilibrium (NAE) for any share of fake reviewers, β, any share of naive consumers, ν, and any number of reviews k. The fake reviewers' reporting strategies are given by*

$$m^F(x) = \begin{cases} x & for \quad x \geq \check{x} \\ 1 - \frac{1-\check{x}}{\check{x}} x & for \quad x < \check{x} \end{cases} \tag{1.8}$$

where $\check{x}$ is in $(0, 1]$.

Proof. See Appendix 1.C.

The first thing to notice about the NAE is that fake reviewers' no longer play mixed strategies in equilibrium, i.e. their strategies are not given by distributions over reviews. Instead, they employ mixed strategies conditioning their reviews on the good's quality. This strategy is truthful for quality levels above the threshold $\check{x}$ and 'inverted' for quality levels below that threshold such that low quality levels are mapped onto high reviews negative assortative. We refer to the threshold $\check{x}$ as *point of deflection*. The NAE is not unique as there exists a NAE for any $\check{x} \in (0, 1]$. In particular, there exists a point of

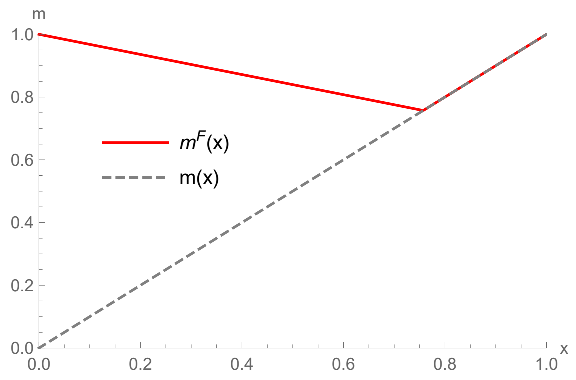


Figure 1.9: Equilibrium reporting strategies in the NAE with $\check{x} = \check{x}_c$.

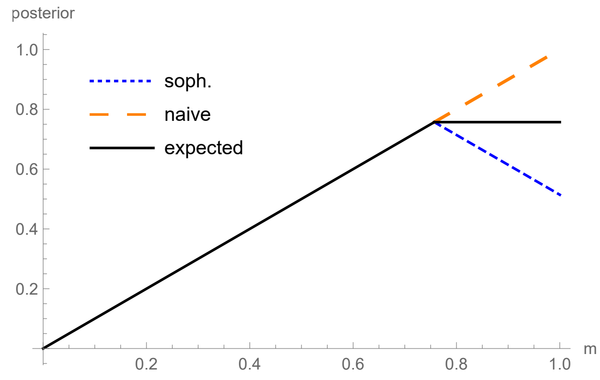


Figure 1.10: Equilibrium posterior beliefs in the NAE with $\check{x} = \check{x}_c$.

deflection $\check{x}_c$ such that the indifference condition in Equation (A.12) is satisfied. The posteriors induced in this equilibrium are equivalent to those induced in the equilibrium of the baseline model for some β.

Corollary 1 *For any β and ν, there exists a point of deflection $\check{x}_c \in (0, 1]$ such that in the corresponding NAE any vector $(m, m, ..., m)$ of reviews with $m \geq \check{x}_c$ induces the same posterior.*
Proof. See Appendix 1.C.

Figures 1.9 and 1.10 illustrate the NAE with deflection point $\check{x}_c \approx 0.76$. In contrast to the previous figures illustrating equilibrium reporting strategies, Figure 1.9 shows the strategies in the (x, m)-plane rather than the $(m, f(m))$-plane. Consequently, the truthtelling strategy is no longer represented by a horizontal line but by the identity line instead.[14] The posterior expectations induced in equilibrium are shown in Figure 1.10 and they coincide with those in the equilibrium of the baseline model.

The extended model with $k \geq 1$ reviewers also has other equilibria. To see this, consider an equilibrium, where both fake and real reviewers report truthfully. No fake reviewer has an incentive to deviate from this equilibrium because if he deviated to a higher message, by $A1$, a consumer would disregard it. If he instead deviated to a lower message, this would lower a consumer's posterior beliefs and therefore reduce the purchasing probability.

As we showed in this section, the baseline model can be extended to allow for multiple consumers and reviewers. While extending it to multiple consumers is straightforward and doesn't change any of our findings, allowing for multiple reviewers complicates the

[14] Note that the fake reviewer's negative assortative reporting strategy does not map onto $f^F(m)$. In the $(m, f(m))$-plane his strategy maps onto a horizontal line at $\frac{1}{1-\check{x}}$.

model and introduces equilibrium multiplicity.

1.8 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 reviewers 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 consumer'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 reviewer:

(d) $m_n^* \in \arg\max_{m\in[0,1]} \mathbb{E}_{Y,T}[a^{T*}(m,Y)x + (1-a_T^*(m,Y))Y]$
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. (2021) full honesty does not arise in equilibrium. Instead, a strategically honest reviewer engages in *underreporting.*

Proposition 5 *A Perfect Bayesian Equilibrium, where all real reviewers tell the truth, does not exist.*

Proof. See Appendix 1.C.

The intuition behind this result is captured by Figure 1.11, 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 1.11 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 $q(m') < m'$, she forgoes the additional utility $m' - y$. Thus, the green area (light + dark) represents the mistake she *avoids* making in expectation. Similarly, the dark green area represents the mistake that a naive consumer makes in

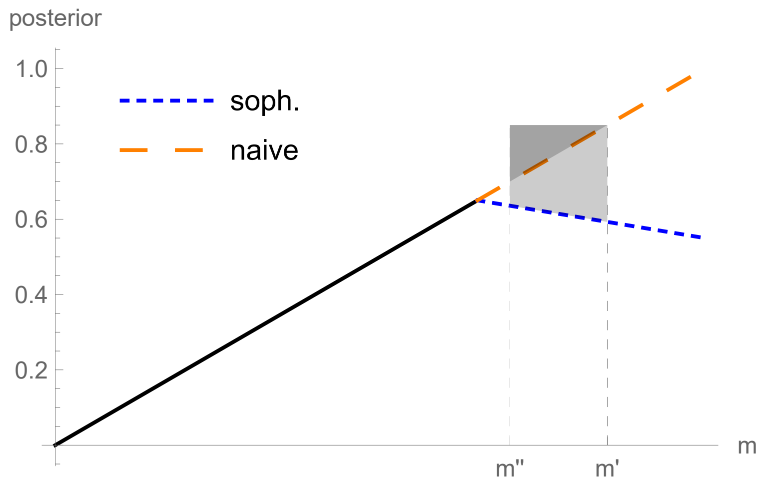


Figure 1.11: Deviation by a strategically honest reviewer

expectation as a result of the strategic deviation. The light green area then illustrates the net benefit of such a strategic deviation. Proposition 6 characterizes the equilibrium of the extended game.

Proposition 6 *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*

$$f^F(m) = (1-\delta)\frac{1-\beta}{\beta}\frac{m-c}{c-(1-\nu)\bar{q}-\nu m} \tag{1.9}$$

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

$$\int_c^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta}{1-\beta}\frac{1-\delta}{\eta} \tag{1.10}$$

Proof. See Appendix 1.C.

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. A fake reviewer 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.

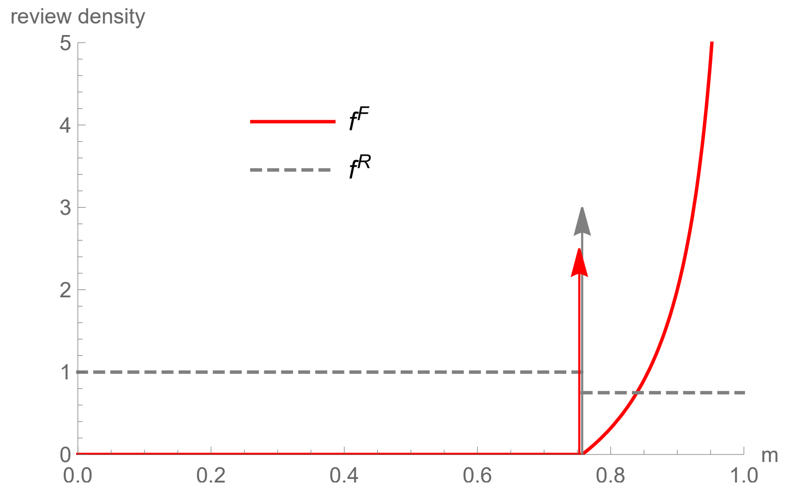


Figure 1.12: Density of real (gray, dashed) and fake (red) reviews in an equilibrium with strategically honest reviewers.

Proposition 7 *If the share of naive consumers is low enough relative to fake reviewers who would exploit them, $\nu < \frac{1-\sqrt{\beta}}{1+\sqrt{\beta}}$, then as honest reviewers become strategic, $\eta \to 0$, the equilibrium outcome is equivalent to eliminating naive consumers, $\nu \to 0$.*
Proof. See Appendix 1.C.

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]

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 7 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

[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 $m = 1$.

platforms are likely easier to reach with educational policies. Proposition 7 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.

1.9 Conclusion

In this paper, I contribute to the ongoing debate about regulating online platforms by theoretically analysing commonly discussed awareness policies. To this end, I propose a theoretical model that captures the key features in this context, fake reviews and naive consumers. I show that awareness policies always trade off the surplus of naive consumers against that of sophisticated ones. However, because the positive effects dominate such policies lead to an increase in aggregate consumer surplus.

Contrary to models without naive consumers, writing honest reviews is not always optimal for real 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 real 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. Awareness policies that target reviewers will therefore also reach consumers and vice versa. If reviewers who become aware of fake reviews start writing reviews strategically, then awareness policies lead to the same outcome in the limit, regardless of whether all consumers, all reviewers, or both, are made aware. Therefore, such policies need not necessarily have a narrow focus but can be implemented broadly.

While the present paper studies the effects of awareness policies it does not directly address the incentives to raise consumer awareness. However, its findings help shed light on this question. First, since sophisticated consumers are harmed by increased awareness, it is unlikely that information about fake reviews disseminates among consumers. Second, a platform that collects fees from sellers is also unlikely to spread this information because fake reviews increase sales and therefore a platform's revenue. This might explain why Amazon highlights its effort to fight fake reviews but does little to inform consumers that despite these efforts they are still prevalent on its website. Sellers who abstain from fake reviews might have such incentives but it might be difficult for them to communicate credibly that their competitors use fake reviews but they do not. While third party services like Fakespot provide information about fake reviews they mainly provide better information to already aware consumers. Thus, the findings in this paper highlight the importance for policymakers to assume an active role in raising consumer awareness.

1.A Calculations for the example in Section 1.2

Example. Given that fake reviews are $5\star$ with prob. q and $4\star$ with prob. $1-q$, the expected product qualities conditional on these claims, according to Bayes' Rule, are

$$\begin{aligned}\mathbb{E}^s[X|5\star] &= Pr(fake|5\star) * \frac{1}{2} + Pr(real|5\star) * 1 \\ &= \frac{q*\frac{1}{2}}{q*\frac{1}{2}+\frac{1}{4}*\frac{1}{2}} * \frac{1}{2} + \frac{\frac{1}{4}*\frac{1}{2}}{q*\frac{1}{2}+\frac{1}{4}*\frac{1}{2}} * 1 = \frac{2q+1}{4q+1} \\ \mathbb{E}^s[X|4\star] &= Pr(fake|4\star) * \frac{1}{2} + Pr(real|4\star) * \frac{3}{4} = \frac{11-8q}{20-16q}.\end{aligned}$$

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 sending a $5\star$ review with probability q and a $4\star$ review with probability $(1-q)$. Using that sophisticates update their beliefs according to Bayes' Rule as above while naives take reviews at face value, we have that the optimal q satisfies:

$$\nu + (1-\nu)\mathbb{E}^s[X|5\star] = \nu\frac{3}{4} + (1-\nu)\mathbb{E}^s[X|4\star]$$

Plugging in $\frac{1}{5}$ for the share of naive consumers, ν, yields $q^* \approx 0.89$.

1.B Proofs of Lemmas and Propositions

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.

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

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 reviewer would find it beneficial to send that message instead. $\square$

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 a 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 $(1-\nu)\bar{q} + \nu$, 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, $(1-\nu)\bar{q} + \nu < 1 - \epsilon \;\Leftrightarrow\; (1-\nu)(1-\bar{q}) > \epsilon$. 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,*

$$c \geq \underline{c} = (1-\nu)\bar{q} + \nu > \bar{q}.$$

Proof. Because the fake reviewer's equilibrium strategy is an atomless distribution f^F, the posterior expectation of a sophisticated consumer in (1.2) can be written as

$$q(m) = \frac{\beta f^F(m)}{1-\beta+\beta f^F(m)}\bar{q} + \frac{1-\beta}{1-\beta+\beta f^F(m)}m,$$

which is a convex combination of $\bar{q}$ and m. The posterior is therefore also a convex combination of $\bar{q}$ and m. Two implications follow from this. Thus, for any review $m < \bar{q}$ we have that $\mathbb{E}_T[\mathbb{E}^T[X|m]] < \bar{q}$, while for any review $m \geq \bar{q}$ we have $\mathbb{E}_T[\mathbb{E}^T[X|m]] \geq \bar{q}$. A fake reviewer can therefore not write reviews below $\bar{q}$ in equilibrium. This means that $supp(f^b) \subseteq [\bar{q}, 1]$. Denote by c the constant posterior that fake reviews induce in equilibrium and notice that reviews above $\bar{q}$ 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 $\underline{c} = (1-\nu)\bar{q} + \nu$, 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. $\square$

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 (1.2) simplifies to:

$$q(m) = \frac{\beta f^F(m)}{1-\beta+\beta f^F(m)}\bar{q} + \frac{(1-\beta)}{1-\beta+\beta f^F(m)}m \tag{A.11}$$

By Lemma 1 we have

$$(1-\nu)q(m) + \nu m = c \tag{A.12}$$

and hence

$$c = (1-\nu)\left(\frac{(1-\beta)}{1-\beta+\beta f^F(m)}m + \frac{\beta f^F(m)}{1-\beta+\beta f^F(m)}\bar{q}\right) + \nu m \tag{A.13}$$

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

$$f^F(m) = \frac{1-\beta}{\beta}\frac{m-c}{c-(1-\nu)\bar{q}-\nu m} \tag{A.14}$$

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:

$$\int_c^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta}{1-\beta} \tag{A.15}$$

What we will establish in this proof is that (A.15) always has a unique solution. For $\beta \in (0, 1)$, the RHS of (A.15) 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

$$\int_c^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m}dm \tag{A.16}$$

can get arbitrarily big. Recall from Lemma 3 the lower bound $\underline{c}=(1-\nu)\bar{q}+\nu$. As $c \to \underline{c}$, we have (A.16) tending to

$$\int_{\underline{c}}^1 \frac{m-(1-\nu)\bar{q}-\nu}{(1-\nu)\bar{q}+\nu-(1-\nu)\bar{q}-\nu m}dm = \int_{\underline{c}}^1 \frac{m-\underline{c}}{\nu(1-m)}dm = \frac{1}{\nu}\int_{\underline{c}}^1 \frac{m-\underline{c}}{1-m}dm$$

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 $g(m)=\frac{\hat{c}-\underline{c}}{1-m}$.

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

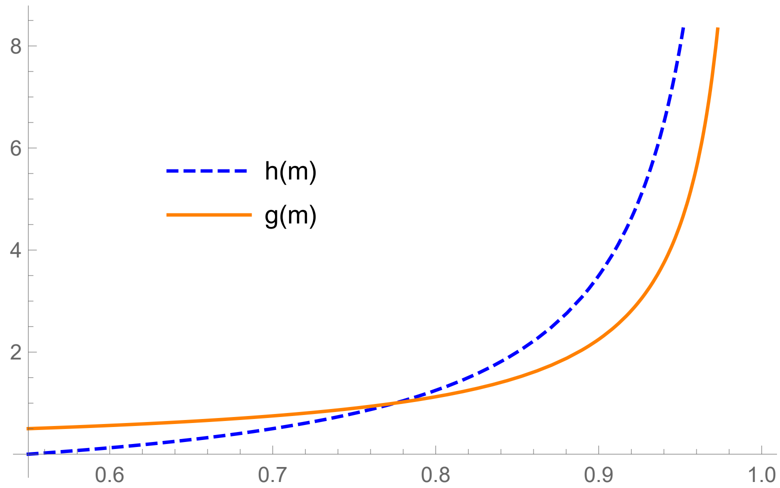


Figure A.13: $h(m)$ in blue vs. $g(m)$ in brown

$\int_{\underline{c}}^1 g(m)dm=\infty$ 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 (A.16) can get arbitrarily small. We can rewrite (A.16), restricting the range of integration using the indicator function, as

$$\int_0^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m}\mathbb{1}_{\{c\le m\le 1\}}dm, \tag{A.17}$$

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

$$\lim_{c\to 1}\int_0^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m}\mathbb{1}_{\{c\le m\le 1\}}dm = \int_0^1 \frac{m-1}{1-(1-\nu)\bar{q}-\nu m}\mathbb{1}_{\{m=1\}}dm.$$

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

$$\lim_{c\to 1}\int_c^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m}dm = 0$$

By the intermediate value theorem, (A.15) must have a solution for any $\beta\in(0,1)$.

Finally, we show that (A.16) is monotonically decreasing in c. Taking the derivative of the left-hand side with respect to c, we have by Leibniz rule:

$$\begin{aligned}\frac{d}{dc}\int_c^1 \frac{(m-c)}{c-(1-\nu)\bar{q}-\nu m}dm &= \int_c^1 \frac{d}{dc}\left(\frac{(m-c)}{c-(1-\nu)\bar{q}-\nu m}\right)dm \\ &= \int_c^1 -\frac{[\overbrace{(c-(1-\nu)\bar{q}-\nu m)}^{\ge 0}+\overbrace{(m-c)}^{\ge 0}]}{\underbrace{(c-(1-\nu)\bar{q}-\nu m)^2}_{\ge 0}}dm\end{aligned}$$

an integral of a function that is negative on the range of integration. Because (A.16) is monotonically decreasing in c, the solution to (A.15) is unique. $\square$

Proof of Proposition 2: To show the first part, note that $\int_c^1 \frac{m-c}{c-(1-\nu)\bar{q}-\nu m}dm$ is increasing in ν and decreasing in c. In equilibrium, as (A.15) 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 $f^F(m)=\frac{1-\beta}{\beta}\frac{m-c}{c-(1-\nu)\bar{q}-\nu m}$ is continuous in $]-\infty,1]$ for $\nu\in[0,1)$. In particular, $c-(1-\nu)\bar{q}-\nu m$ 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. $\lim_{m\to-\infty} f^F(m)$, by L'Hospital's rule, is equal to $-\frac{1}{\nu}$ and, hence, $\lim_{m\to-\infty} f^{F,\nu_2}(m) > \lim_{m\to-\infty} f^{F,\nu_1}(m)$. 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:
We first prove a useful Lemma.
Lemma 4. *Let $g(x)$ be a continous function with $g(x) > 0$ for $x \in [a, b]$. Let $h(x)$ be a continuous and monotonically increasing function with $\int_a^b h(x) = 0$.*
If g is strictly increasing (decreasing) on $[a, b]$, then $\int_a^b g(x)h(x)dx > 0$ (< 0).
Proof: We prove the case for increasing g. The case for decreasing g can be proved by simply reversing all inequalities.
$h(x)$ continuous and monotonically increasing and $\int_a^b h(x) = 0$ implies that $\exists \tilde{x} \in (a, b)$ s.t. (i) $h(\tilde{x}) = 0$, (ii) $h(x) < 0 \forall x \in [a, \tilde{x})$, (iii) $h(x) > 0 \forall x \in [\tilde{x}, b)$, and (iv) -$\int_a^{\tilde{x}} h(x)\, dx = \int_{\tilde{x}}^b h(x)\, dx$. We can rewrite $\int_a^b g(x)h(x)\, dx$ as $\int_a^{\tilde{x}} g(x)h(x)\, dx + \int_{\tilde{x}}^b g(x)h(x)\, dx$ and notice that, since g is strictly increasing, $g(a) \int_a^{\tilde{x}} h(x)\, dx > \int_a^{\tilde{x}} g(x)h(x)\, dx > g(\tilde{x}) \int_a^{\tilde{x}} h(x)\, dx$. By the IVT, $\exists \underline{x} \in (a, \tilde{x})$ s.t. $g(\underline{x}) \int_a^{\tilde{x}} h(x)\, dx = \int_a^{\tilde{x}} g(x)h(x)\, dx$. Equivalently, we have $g(\tilde{x}) \int_{\tilde{x}}^b h(x)\, dx < \int_{\tilde{x}}^b g(x)h(x)dx < g(b) \int_{\tilde{x}}^b h(x)\, dx$ and $\exists \bar{x} \in (\tilde{x}, b)$ s.t. $g(\bar{x}) \int_{\tilde{x}}^b h(x)\, dx = \int_{\tilde{x}}^b g(x)h(x)\, dx$. As strictly increasing g implies $g(\underline{x}) < g(\bar{x})$ and due to (iv) we have $-g(\underline{x}) \int_a^{\tilde{x}} h(x)\, dx < g(\bar{x}) \int_{\tilde{x}}^b g(x)h(x)\, dx \implies \int_a^{\tilde{x}} g(x)h(x)\, dx + \int_{\tilde{x}}^b h(x)\, dx > 0$. □
We now turn to the proof of Proposition 3.
part (i): The consumer surplus of a naive consumer is given by

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

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

$$CS^n = \frac{\beta}{2} \int_c^1 \left(1 - m^2 + m\right) f^F(m)\, dm + \frac{(1-\beta)}{2} \int_0^1 1 + m^2\, dm. \tag{A.18}$$

The second integral is independent of ν and thus, by Leibniz Rule, we have

$$\frac{dCS^n}{d\nu} = \frac{\beta}{2} \int_c^1 \left(1 - m^2 + m\right) \frac{df^F(m)}{d\nu}\, dm \tag{A.19}$$

$1 - m^2 + m$ in strictly decreasing for $m \in [\frac{1}{2}, 1]$ and, since $c > \frac{1}{2}$, also for $m \in [c, 1]$. By Lemma 4 we then have that $\frac{dCS^n}{d\nu} < 0$.

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

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

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

$$\begin{aligned} CS^s =& \frac{\beta}{2} \int_c^1 \left(1 - q(m)^2 + 2q(m)\bar{q}\right)\ f^F(m)\ dm \\ &+ \frac{(1-\beta)}{2} \int_0^c 1 + m^2\ \ dm \\ &+ \frac{(1-\beta)}{2} \int_c^1 1 - q(m)^2 + 2q(m)m\ dm \end{aligned} \tag{A.20}$$

The derivative with respect to ν, using Leibniz Rule and rearranging, is

$$\begin{aligned} \frac{dCS^s}{d\nu} =& \left(-\frac{\beta}{2}\left(1 - q(c)^2 + 2q(c)\bar{q}\right) f^F(c) + \frac{1-\beta}{2}\left((1 + c^2) - (1 + c^2)\right)\right) \frac{dc}{d\nu} \\ &+ \beta \int_c^1 \left(\bar{q} - q(m)\right) f^F(m) \frac{dq(m)}{d\nu}\, dm + (1-\beta) \int_c^1 \left(m - q(m)\right) \frac{dq(m)}{d\nu}\, dm \\ &+ \frac{\beta}{2} \int_c^1 \left(1 - q(m)^2 + 2q(m)\bar{q}\right) \frac{df^F(m)}{d\nu}\, dm \end{aligned}$$

Because $f^F(c) = 0$, it is easy to see that the term on the first line is zero. The term on the second line can be rewritten as

$$\int_c^1 \left(\beta\bar{q}f^F(m) + (1-\beta)m - \left(\beta f^F(m) + (1-\beta)\right)q(m)\right) \frac{dq(m)}{d\nu}\, dm$$

and because $q(m) = \frac{\beta f^F(m)\bar{q} + (1-\beta)m}{1-\beta+\beta f^F(m)}$ the term on the second line is also zero. The term on the third line contains an expression of the form $\int_c^1 g(m)h(m)dm$, where $g(m) = 1 - q(m)^2 + 2q(m)\bar{q}$ is strictly increasing in m for $m \in [c, 1]$ and $h(m) = \frac{df^F(m)}{d\nu}$ satisfies $\int_c^1 h(x) = 0$. Therefore, by Lemma 4, the term on the third line is positive and we have $\frac{dCS^s}{d\nu} > 0$.

part (iii):

We will prove that the marginal effect of informing consumers is positive in the limits as $\nu \to 0$ and $\nu \to 1$ and show numerically that this holds also for intermediate values of ν. Total consumer surplus is given by

$$CS = \nu\, CS^n + (1-\nu)\, CS^s \tag{A.21}$$

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

$$\frac{dCS}{d\nu} = \underbrace{(CS^n - CS^s)}_{A} + \underbrace{\nu \frac{dCS^n}{d\nu}}_{B} + \underbrace{(1-\nu)\frac{dCS^s}{d\nu}}_{C} \tag{A.22}$$

Note that $\frac{dCS}{d\nu}$ reflects an increase in the share of naive consumers which is the opposite of an educational policy. We thus need to show that Equation (A.22) is negative. We begin by showing that A is negative. To do so, it suffices to show it in the limit as $\nu \to 0$. This is because of the fact that in parts (i) and (ii) we showed that $\frac{dCS^n}{d\nu} < 0$ and $\frac{dCS^s}{d\nu} > 0$. From Equations (A.18) and (A.20) we write A as:

$$\begin{aligned} CS^n - CS^s =& \frac{\beta}{2}\int_c^1 \left((m-m^2) - (q(m) - q(m)^2)\right) f^F(m)\, dm \\ &- \frac{1-\beta}{2}\int_c^1 1 + 2q(m)m - q(m)^2\, dm \end{aligned} \tag{A.23}$$

From Eq. (A.12) we have that $q(m) = \frac{c-\nu m}{1-\nu}$ and therefore $\lim\limits_{\nu\to 0} q(m) = c$. Furthermore, we have $\lim\limits_{\nu\to 0} f^F(m) = \lim\limits_{\nu\to 0} \frac{1-\beta}{\beta}\frac{m-c}{c-\frac{1-\nu}{2}-\nu m} = \frac{1-\beta}{\beta}\frac{2(m-c)}{2c-1}$ and therefore:

$$\begin{aligned} \lim_{\nu\to 0}(CS^s - CS^n) &= \frac{\beta}{2}\int_c^1 \frac{1-\beta}{\beta}\left((m-c) - (m^2-c^2)\right)\frac{2(m-c)}{2c-1}\, dm - \frac{1-\beta}{2}\int_c^1 1 + 2cm - c^2\, dm \\ &= \frac{1}{(1-\beta)(2c-1)}\int_c^1 (1-(m+c))\,(m-c)^2\, dm - \frac{1-\beta}{2}\int_c^1 1 + 2cm - c^2\, dm \end{aligned}$$

Because $m+c > 1$ for $m \in [c,1]$, the first integral is negative and because $c^2 < 1$, the second one is positive. Therefore, $\lim\limits_{\nu\to 0}(CS^s - CS^n) < 0$ and $A > 0$ holds $\forall \nu \in (0,1)$. Next we show that $B = 0$ in both limiting cases. Let us consider first $\nu \to 0$, in which case $c \to c^{JO} = \frac{1}{1+\sqrt{\beta}}$. From Eq. (A.19) we then have $-\infty < \lim\limits_{\nu\to 0}\frac{dCS^n}{d\nu} = \frac{\beta}{2}\int_{c^{JO}}^1 (1-m^2+m)\frac{df^F(m)}{d\nu}\, dm < 0$ and thus $\lim\limits_{\nu\to 0}\nu\frac{dCS^n}{d\nu} = 0$. In the limiting case $\nu \to 1$ we have that $c \to 1$ and therefore $\lim\limits_{\nu\to 1}\frac{dCS^n}{d\nu} = 0$, which in turn implies $\lim\limits_{\nu\to 1}\nu\frac{dCS^n}{d\nu} = 0$. Hence, $\lim\limits_{\nu\to 0} B = \lim\limits_{\nu\to 1} B = 0$. What is left to show is that $C = 0$ in both limiting cases. Beginning again with $\nu \to 0$ and using again $\lim\limits_{\nu\to 0} q(m) = c$ we obtain $\lim\limits_{\nu\to 0}\frac{dCS^s}{d\nu} = (1+c-c^2)\int_{c^{JO}}^1 \frac{df^F(m)}{d\nu}\, dm = 0$ and thus $\lim\limits_{\nu\to 0}(1-\nu)\frac{dCS^s}{d\nu}=0$. In the limiting case $\nu\to 1$, the fact that $c\to 1$ again gives us $\lim\limits_{\nu\to 1}\frac{dCS^s}{d\nu}=0$ and consequently, $\lim\limits_{\nu\to 1}(1-\nu)\frac{dCS^n}{d\nu}=0$. Hence, $\lim\limits_{\nu\to 0}C=\lim\limits_{\nu\to 1}C=0$. Combining the results for A, B, and C, we finally have that Hence, $\lim\limits_{\nu\to 0}\frac{dCS}{d\nu}<0$ and $\lim\limits_{\nu\to 1}\frac{dCS}{d\nu}<0$. □

1.C Proofs for extensions

Proof of Proposition 1b: We first prove Lemma 5, which corresponds to Lemmas $1-3$ of the baseline model.

Lemma 5. *In equilibrium, the two fake reviewer types mix over disjoint sets of messages. Their equilibrium strategies are represented by atomless distributions f_H^F and f_L^F. Furthermore, $supp(f_H^F)=[c_H,1]$ and $supp(f_L^F)=[0,c_L]$ with $c_L<\bar{q}<c_H$.*

Proof: First note that a negative fake reviewer, in his effort to minimize the purchasing probability, induces the lowest posterior with all messages that he sends in equilibrium. The argument is similar to the one in Lemma 1, but instead of maximizing it, the negative fake reviewers are minimizing the posterior. Next, we show that there is no message that both types send. The posterior induced by any message m, $\mathbb{E}_T[\mathbb{E}^T[X|m]]$ is a convex combination of $\bar{q}$ and m. Whenever $\nu>0$, $\mathbb{E}_T[\mathbb{E}^T[X|m]]<\bar{q}$ for any $m<\bar{q}$ and $\mathbb{E}_T[\mathbb{E}^T[X|m]]>\bar{q}$ for any $m>\bar{q}$. Then positive (negative) types can only send messages strictly above (below) $\bar{q}$. When $\nu=0$, then $\mathbb{E}_T[\mathbb{E}^T[X|m]]=q(m)=\bar{q}$ whenever a fake reviewer sends m with a positive probability. But since there can only be countably many such atoms, there are messages below and above $\bar{q}$ that are not sent with positive probability and therefore induce strictly lower and strictly higher posteriors. This implies that all messages that a positive (negative) type sends must induce posteriors strictly above (below) $\bar{q}$ and thus the messages themselves are strictly above (below) $\bar{q}$. This proves the first statement of the lemma.

The proof of the second statement is analogous to the corresponding statement in the baseline model (without the negative type) and we therefore refer to the proof of the second part of Lemma 3. □

We now turn to the proof of Proposition 1*b*. For the positive type, the proof is the same as that of Proposition 1 except for replacing c with c_H and β with β_H. What we show now is that an equivalent proof goes through also for the negative type. First, to simplify notation, we define $\beta:=\beta_L+\beta_H$. Because the fake reviewers' equilibrium strategies are atomless distributions we have:

$$q(m)=\frac{\beta_L f_L^F(m)+\beta_H f_H^F(m)}{1-\beta+\beta_L f_L^F(m)+\beta_H f_H^F(m)}\bar{q}+\frac{(1-\beta)}{1-\beta+\beta_L f_L^F(m)+\beta_H f_H^F(m)}m. \tag{A.24}$$

Note that $f_H^F(m)=0$ for $m\in[0,c_H]$ and $f_L^F(m)=0$ for $m\in[c_L,1]$ which reduces the sophisticated consumer's posterior expectation to

$$q(m) = \frac{\beta_L f_L^F(m)}{1-\beta+\beta_L f_L^F(m)}\bar{q} + \frac{(1-\beta)}{1-\beta+\beta_L f_L^F(m)}m. \tag{A.25}$$

and

$$q(m) = \frac{\beta_H f_H^F(m)}{1-\beta+\beta_H f_H^F(m)}\bar{q} + \frac{(1-\beta)}{1-\beta+\beta_H f_H^F(m)}m. \tag{A.26}$$

in these cases. We will use the reduced expressions whenever these cases apply to simplify notation. Combining this with $(1-\nu)q(m)+\nu m = c_L$ allows us to solve for f_L^F.

$$f_L^F(m) = \frac{1-\beta}{\beta_L}\frac{m-c_L}{c_L-(1-\nu)\bar{q}-\nu m} \tag{A.27}$$

which has to integrate to 1 over its support $[0, c_L]$. Therefore, the following equation implicitly defines the equilibrium value c_L.

$$\int_0^{c_L} \frac{m-c_L}{c_L-(1-\nu)\bar{q}-\nu m}dm = \frac{\beta_L}{1-\beta} \tag{A.28}$$

To show that a solution to Equation (A.28) always exists we begin by showing that the LHS can get arbitrarily big setting c_L equal to its upper bound $\underline{c}_L = (1-\nu)\bar{q}$, which is the highest expectation that sending $m = 0$ can induce. In this case, the LHS can be simplified to

$$\frac{1}{\nu}\int_0^{c_L} \frac{\underline{c}_L - m}{m}dm. \tag{A.29}$$

Because $\lim\limits_{c' \to \underline{c}_L} \int_0^{c'} \frac{\underline{c}_L - m}{m}dm = \infty$, the LHS is not bounded from above. To show that it can get arbitrarily small rewrite the LHS of Equation (A.28) as

$$\int_{-\infty}^{\infty} \frac{m-c_L}{c_L-(1-\nu)\bar{q}-\nu m}\mathbb{1}_{\{0\le m\le \underline{c}_L\}}dm \tag{A.30}$$

and consider the limit as $\underline{c}_L \to 0$.

$$\lim_{\underline{c}_L\to 0}\int_{-\infty}^{\infty} \frac{m-c_L}{c_L-(1-\nu)\bar{q}-\nu m}\mathbb{1}_{\{0\le m\le \underline{c}_L\}}dm = \int_{-\infty}^{\infty} \frac{m}{(1-\nu)\bar{q}-\nu m}\mathbb{1}_{\{m=0\}}dm \tag{A.31}$$

For $m = 0$ we have $\frac{m}{(1-\nu)\bar{q}-\nu m} = 0$ while for $m \neq 0$ we have $\mathbb{1}_{\{m=0\}} = 0$ and thus the limit of the LHS as $c_L \to 0$ is 0. Since the LHS of Equation (A.28) is a continuous function of c_L, the intermediate value theorem guarantees that a solution for Equation (A.28) always exists. What remains to be shown is that this solution is always unique. For this we take the derivative of the LHS of Equation (A.28) applying Leinbitz' Rule:

$$\frac{d}{dc_L}\int_0^{c_L} \frac{(m-c_L)}{c_L-(1-\nu)\bar{q}-\nu m}dm = \int_0^{c_L} \frac{d}{dc_L}\left(\frac{(m-c_L)}{c_L-(1-\nu)\bar{q}-\nu m}\right)dm$$

$$= \int_0^{c_L} -\frac{[\overbrace{(c_L-(1-\nu)\bar{q}-\nu m)}^{<0} + \overbrace{(m-c_L)}^{\le 0}]}{\underbrace{(c_L-(1-\nu)\bar{q}-\nu m)^2}_{>0}}dm$$

We have the integral of a function that is positive on the entire range of integration and thus the LHS of Equation (A.28) is monotonically increasing in c_L between 0 and c_L, which implies that the solution to Equation (A.28) is unique. $\square$

Proof of Proposition 3b: We will utilize Lemma 4b, which is a slight modification of Lemma 4.

Lemma 4b. *Let $g(x)$ be a continous function with $g(x) > 0$ for $x \in [a,b]$. Let $h(x)$ be a continuous and monotonically decreasing function with $\int_a^b h(x) = 0$.*
If g is strictly increasing (decreasing) on $[a,b]$, then $\int_a^b g(x)h(x)dx < 0$ (> 0).

Proof: The only difference to Lemma 4 is that $h(x)$ is decreasing instead of increasing. Changing the inequality signs accordingly provides the proof. $\square$

We now turn to the proof of Proposition 3b.

part (i): The consumer surplus of a naive consumer is given by

$$\begin{aligned} CS^n =& \beta_H \int_{c_H}^{1} \left([1 - F_Y(m)]\ \mathbb{E}[Y|Y > m] + F_Y(m)\ \bar{q} \right) f_H^F(m)\, dm \\ &+ \beta_L \int_{0}^{c_L} \left([1 - F_Y(m)]\ \mathbb{E}[Y|Y > m] + F_Y(m)\ \bar{q} \right) f_L^F(m)\, dm \\ &+ (1-\beta) \int_{0}^{1} \left([1 - F_Y(m)]\ \mathbb{E}[Y|Y > m] + F_Y(m)\ m \right)\, dm, \end{aligned} \tag{A.32}$$

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

$$\begin{aligned} CS^n =& \frac{\beta_H}{2} \int_{c_H}^{1} \left(1 - m^2 + m\right) f_H^F(m)\, dm + \frac{\beta_L}{2} \int_{0}^{c_L} \left(1 - m^2 + m\right) f_L^F(m)\, dm \\ &+ \frac{(1-\beta)}{2} \int_{0}^{1} 1 + m^2\, dm. \end{aligned} \tag{A.33}$$

The last integral is independent of ν and thus, by Leibniz' Rule, using the fact that $f_H^F(c_H) = 0$ and $f_L^F(c_L) = 0$ we have

$$\frac{dCS^n}{d\nu} = \frac{\beta_H}{2} \int_{c_H}^{1} \left(1 - m^2 + m\right) \frac{df_H^F(m)}{d\nu}\, dm + \frac{\beta_L}{2} \int_{0}^{c_L} \left(1 - m^2 + m\right) \frac{df_L^F(m)}{d\nu}\, dm. \tag{A.34}$$

Note that $1 - m^2 + m$ is strictly increasing for $m \in [0, \frac{1}{2})$ and strictly decreasing for $m \in (\frac{1}{2}, 1]$. Since $c_H > \frac{1}{2}$ and $c_L < \frac{1}{2}$, this is also true for $m \in [c_H, 1]$ and $m \in [0, c_L]$. By Lemma 4 the first integral is negative and by Lemma 4b the second one is negative. $\square$

part (ii): A sophisticated consumer's ex ante expected surplus is given by

$$\begin{aligned}CS^s =& \beta_H \int_{c_H}^{1} \left([1-F_Y(q(m))]\ \mathbb{E}[Y|Y>q(m)] + F_Y(q(m))\ \bar{q}\right) f_H^F(m)\, dm \\ &+ \beta_L \int_{0}^{c_L} \left([1-F_Y(q(m))]\ \mathbb{E}[Y|Y>q(m)] + F_Y(q(m))\ \bar{q}\right) f_L^F(m)\, dm \\ &+ (1-\beta) \int_{0}^{1} \left([1-F_Y(q(m))]\ \mathbb{E}[Y|Y>q(m)] + F_Y(q(m))\ m\right)\, dm,\end{aligned}$$

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

$$\begin{aligned}CS^s =& \frac{\beta_H}{2} \int_{c_H}^{1} \left(1-q(m)^2+2q(m)\bar{q}\right)\ f_H^F(m)\ dm \\ &+ \frac{\beta_L}{2} \int_{0}^{c_L} \left(1-q(m)^2+2q(m)\bar{q}\right)\ f_L^F(m)\ dm \\ &+ \frac{(1-\beta)}{2} \int_{0}^{1} \left(1-q(m)^2+2q(m)m\right)\ \ dm\end{aligned} \tag{A.35}$$

The derivative with respect to ν, using using the fact that $f_H^F(c_H)=0$ and $f_L^F(c_L)=0$ and rearranging, is

$$\begin{aligned}\frac{dCS^s}{d\nu} =& \frac{\beta_H}{2} \int_{c_H}^{1} \Big(1-q(m)^2+2q(m)\bar{q}\Big) \frac{df_H^F(m)}{d\nu}\, dm + \frac{\beta_L}{2} \int_{0}^{c_L} \Big(1-q(m)^2+2q(m)\bar{q}\Big) \frac{df_L^F(m)}{d\nu}\, dm \\ &+ \beta_H \int_{c_H}^{1} \Big(\bar{q}-q(m)\Big) f_H^F(m) \frac{dq(m)}{d\nu}\, dm + (1-\beta) \int_{c_H}^{1} \Big(m-q(m)\Big) \frac{dq(m)}{d\nu}\, dm \\ &+ \beta_L \int_{0}^{c_L} \Big(\bar{q}-q(m)\Big) f_L^F(m) \frac{dq(m)}{d\nu}\, dm + (1-\beta) \int_{0}^{c_H} \Big(m-q(m)\Big) \frac{dq(m)}{d\nu}\, dm.\end{aligned} \tag{A.36}$$

The two integrals on the first line of Equation (A.36) are positive by Lemmas 4 and 4b, respectively. The remaining terms can be rewritten as

$$\begin{aligned}\frac{dCS^s}{d\nu} =& \int_{c_H}^{1} \Big[\beta_H \bar{q} f_H^F(m) + (1-\beta)m - \Big(\beta_H f_H^F(m) + (1-\beta)\Big) q(m)\Big] \frac{dq(m)}{d\nu}\, dm \\ &+ \int_{0}^{c_L} \Big[\beta_L \bar{q} f_L^F(m) + (1-\beta)m - \Big(\beta_L f_L^F(m) + (1-\beta)\Big) q(m)\Big] \frac{dq(m)}{d\nu}\, dm\end{aligned} \tag{A.37}$$

and shown to be 0 using $q(m) = \frac{\beta_i \bar{q} f_i^F(m) + (1-\beta)m}{\beta_i f_i^f(m) + 1 - \beta}$. □

part (iii): We will prove that the marginal effect of informational policies is positive in the limiting cases of $\nu \to 0$ and $\nu \to 1$ and rely on numerical results for intermediate values of ν. The marginal effect is given by

$$\frac{dCS}{d\nu} = \underbrace{(CS^n - CS^s)}_{A} + \underbrace{\nu \frac{dCS^n}{d\nu}}_{B} + \underbrace{(1-\nu)\frac{dCS^s}{d\nu}}_{C}. \tag{A.38}$$

We begin by showing that A is negative. Because $\frac{dCS^n}{d\nu} < 0$ and $\frac{dCS^s}{d\nu} > 0$, it suffices to show it in the limiting case of $\nu \to 0$. From Equations (A.33) and (A.36) we write A as:

$$\begin{aligned} CS^n - CS^s =& \frac{\beta_L}{2} \int_0^{c_L} \Big((m - m^2) - (q(m) - q(m)^2) \Big) f_L^F(m)\, dm \\ &+ \frac{\beta_H}{2} \int_{c_H}^1 \Big((m - m^2) - (q(m) - q(m)^2) \Big) f_H^F(m)\, dm \\ &+ \frac{1-\beta}{2} \left(\int_0^{c_L} (m - q(m))^2\, dm + \int_{c_H}^1 (m - q(m))^2\, dm \right) \end{aligned} \tag{A.39}$$

Using $\lim\limits_{\nu \to 0} q(m) = c_i$ and $\lim\limits_{\nu \to 0} f_i^F(m) = \frac{1-\beta}{\beta_i} \frac{m - c_i}{c_i - \bar{q}}$ for $i = L, H$ it can be shown that

$$\begin{aligned} \lim_{\nu \to 0} CS^n - CS^s =& 2(1-\beta) \int_0^{c_L} (1 - m - c_L) \frac{(m - c_L)^2}{c_L - \bar{q}}\, dm \\ &+ 2(1-\beta) \int_{c_H}^1 (1 - m - c_H) \frac{(m - c_H)^2}{c_H - \bar{q}}\, dm \end{aligned} \tag{A.40}$$

The first integral is negative because $c_L < \bar{q}$ and $m + c_L < 1$ for $m \in [0, c_L]$, while the second integral is negative because $c_H > \bar{q}$ and $m + c_H > 1$ for $m \in [0, c_H]$. Consequently, $A < 0$. We continue by showing that $B = 0$ in both limiting cases. Consider first the case $\nu \to 0$, in which case we have $-\infty < \lim\limits_{\nu \to 0} \frac{dCS^n}{d\nu} = \frac{\beta_H}{2} \int_{c_H}^1 (1 - m^2 + m) \frac{df_H^F(m)}{d\nu}\, dm + \frac{\beta_L}{2} \int_0^{c_L} (1 - m^2 + m) \frac{df_L^F(m)}{d\nu}\, dm < 0$ and therefore $\lim\limits_{\nu \to 0} \nu \frac{dCS^n}{d\nu} = 0$. In the case $\nu \to 1$, we have $\lim\limits_{\nu \to 1} c_H = 1$ and $\lim\limits_{\nu \to 1} c_L = 0$ and therefore $\lim\limits_{\nu \to 1} \frac{dCS^n}{d\nu} = 0$, which in turn implies $\lim\limits_{\nu \to 1} \nu \frac{dCS^n}{d\nu} = 0$. Hence, $\lim\limits_{\nu \to 0} B = \lim\limits_{\nu \to 1} B = 0$. Finally, we show that $C = 0$ in both limiting cases by first considering the case $\nu \to 0$. Using $\lim\limits_{\nu \to 0} q(m) = c_L$ for $m \in [0, c_L]$ and $\lim\limits_{\nu \to 0} q(m) = c_H$ for $m \in [c_H, 1]$ we obtain $\lim\limits_{\nu \to 0} \frac{dCS^s}{d\nu} = \frac{\beta_H}{2}(1 + c_H - c_H^2) \int_{c_H}^1 \frac{df_H^F(m)}{d\nu}\, dm + \frac{\beta_L}{2}(1 + c_L - c_L^2) \int_0^{c_L} \frac{df_L^F(m)}{d\nu}\, dm = 0$ and therefore $\lim\limits_{\nu \to 0}(1 - \nu) \frac{dCS^s}{d\nu} = 0$. In the limiting case $\nu \to 1$ we again have $\lim\limits_{\nu \to 1} c_H = 1$ and $\lim\limits_{\nu \to 1} c_L = 0$, which implies $\lim\limits_{\nu \to 1} \frac{dCS^s}{d\nu} = 0$ and consequently $\lim\limits_{\nu \to 1}(1 - \nu) \frac{dCS^s}{d\nu} = 0$. Hence, $\lim\limits_{\nu \to 0} C = \lim\limits_{\nu \to 1} C = 0$. □

Proof of Proposition 4: In an NAE, at most two different reviews are sent. When the good's quality x is above $\check{x}$ both review types report truthfully, while fake reviewers send $1 - \frac{1-\check{x}}{\check{x}} x$ when $x < \check{x}$. We show that in both cases, fake reviewers have no incentive to deviate. First, if $x > \check{x}$ the only on-schedule deviation is inverting the equilibrium reporting strategy and sending $m^{F,-1}(x) = \frac{\check{x}(1-x)}{1-\check{x}}$. But this leads both consumer types to update their beliefs to $\frac{\check{x}(1-x)}{1-\check{x}} < x$ which lowers the purchasing probability. An off-schedule deviation review m' can be either above or below x. If it is below, then by $A1$ this induces beliefs of $m' < x$ which again reduces the purchasing probability. If it is above x, then any of the other reviews that a consumer sees are below the deviation review, and by $A1$

the deviation has no effect on posterior beliefs. Second, if $x \leq \check{x}$, the only off-schedule deviation is to again invert the reporting strategy which results in a truthful review. If a real review is in the set of reviews that the consumer sees this has no effect while it lowers their posterior beliefs if only fake reviews are seen. For off-schedule deviations, the same argument as before applies and therefore fake reviewers have no incentives to deviate. □

Proof of Corollary 1: We start by examining consumers' posterior beliefs in equilibrium. First, naive consumers trust a set of reviews only if all of them coincide and otherwise update beliefs according to $A1$. Second, sophisticated consumers use Bayes' Rule. Whenever they observe different reviews - because some are truthful and others fake - Bayes' Rule implies that the quality is equal to the lower review. Thus, such a set of reviews is revealing the good's quality. The same is true in the case when all reviews coincide and are below $\check{x}$. Things get more involved when consumers observe only reviews above $\check{x}$ because they could all be fake or at least one could be truthful. Their posterior expectation, given a vector of reviews $\mathbf{m} = (m, m, ..., m)$, is then given by

$$
\begin{aligned}
q(\mathbf{m}) &= Pr(X = m|\mathbf{m})m + Pr(X = m^{F,-1}(m)|\mathbf{m})m^{F,-1}(x) \\
&= \Bigg(\underbrace{Pr(X = m|\mathbf{m}, \neg D)}_{1}\underbrace{Pr(\neg D|\mathbf{m})}_{A} + \underbrace{Pr(X = m|\mathbf{m}, D)}_{B}\underbrace{Pr(D|\mathbf{m})}_{-A}\Bigg)m \qquad \text{(A.41)} \\
&+ \Bigg(\underbrace{Pr(X = m^{F,-1}(m)|\mathbf{m}, \neg D)}_{0}\underbrace{Pr(\neg D|\mathbf{m})}_{A} + \underbrace{Pr(X = m^{F,-1}(m)|\mathbf{m}, D)}_{-B}\underbrace{Pr(D|\mathbf{m})}_{-A}\Bigg)m^{F,-1}(x)
\end{aligned}
$$

We denote by D the event in which all reviews are fake and hence with $\neg D$ the event in which at least one review is real. The probabilities A and B are obtained via Bayes' Rule. For A we have

$$
\begin{aligned}
Pr(\neg D|\mathbf{m}) =& \frac{f(\mathbf{m}|\neg D)Pr(\neg D)}{f(\mathbf{m}|\neg D)Pr(\neg D) + f(\mathbf{m}|D)Pr(D)} \\
&+ \frac{f(X = m|\neg D)Pr(\neg D)}{f(X = m|\neg D)Pr(\neg D) + f(\mathbf{m}|D)Pr(D)} \qquad \text{(A.42)}
\end{aligned}
$$

Because X is independent of D, we have $f(X = m|\neg D) = f(X = m) = 1$, while $Pr(D) = 1 - Pr(\neg D) = \beta^k$. The density $f(\mathbf{m}|D)$ can be derived by looking first at the cdf $F(\mathbf{m}|D) = Pr(\mathbf{m} \leq x|D) = Pr(m^{F,-1}(x) \leq X \leq x) = F(x) - F(m^{F,-1}(x)) = 1 + \frac{\check{x}}{1-\check{x}}$ and taking its derivative $\frac{d}{dx}F(\mathbf{m}|D) = f(\mathbf{m}|D) = \frac{1}{1-\check{x}}$. We can thus write

$$
Pr(\neg D|\mathbf{m}) = \frac{1 - \beta^k}{1 - \beta^k + \frac{\beta^k}{1-\check{x}}} = \frac{(1 - \beta^k)(1 - \check{x})}{(1 - \beta^k)(1 - \check{x}) + \beta^k} \qquad \text{(A.43)}
$$

For the term B we have

$$Pr(X = m|\mathbf{m}, D) = \frac{f(\mathbf{m}|X = m, D)f(X = m|D)}{f(\mathbf{m}|D)} \tag{A.44}$$

which, using again the fact that X does not depend on D and thus $f(X = m|D) = f(X = m) = 1$, that $f(\mathbf{m}|\mathbf{X} = \mathbf{m}, \mathbf{D}) = 1$, and $f(\mathbf{m}|D) = \frac{1}{1-\check{x}}$, we can rewrite this as

$$Pr(X = m|\mathbf{m}, D) = 1 - \check{x} \tag{A.45}$$

and finally, plugging the obtained expressions for A and B into Equation (1.C) and rearranging, we get

$$q(m) = \frac{(1-\check{x})^2 - \beta^k \check{x}^2}{(1-\beta^k)(1-\check{x})^2 + \beta^k(1-\check{x})} m + \frac{\beta^k \check{x}^2}{(1-\beta^k)(1-\check{x})^2 + \beta^k(1-\check{x})} \tag{A.46}$$

The sophisticates' posterior in an NAE is thus linear in m with a slope that is decreasing in $\check{x}$. In order to satisfy the indifference condition in Equation (A.12), we need $\frac{d}{dm}q(m) = \frac{\nu}{\nu-1} < 0$. From Equation (A.46) we can see that $\lim\limits_{\check{x}\to 0} = \beta^k > 0$ and $\lim\limits_{\check{x}\to 1} = -\infty$. Because the slope is a continuous function of $\check{x}$ we have by the IVT that $\exists \check{c} \in (0,1)$ such that $\frac{d}{dm}q(m) = \frac{\nu}{\nu-1}$. □

Proof of Proposition 5: 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 takes 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. 1.4). 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 $q(m) = \frac{c-\nu m}{1-\nu} < m$. Thus, whenever their outside option is between $q(m)$ 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 $q(m)$ and $q(m')$. Thus her expected *avoided* mistake is given by

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

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:

$$\nu\left(m\xi - \frac{\xi^2}{2}\right)\Big|_{c}^{m} < (1-\nu)\left(m\xi - \frac{\xi^2}{2}\right)\Big|_{\xi=q(m)}^{\xi=q(c)}$$

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

Note that $q(m) = \frac{c-\nu m}{1-\nu}$ and thus in particular $q(c) = c$. 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 6: 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 5, 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

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

Two things need to be true. First, $q(c) = c$ so that strategically honest reviewers do not have an incentive to deviate when $x = c$. For $q(c) \neq c$ 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 $q(m) = c$ for all $m \in [c,1]$. The first fact yields

$$\frac{\delta\beta}{\delta\beta + (1-c)(1-\beta)(1-\eta)}\frac{1}{2} + \frac{(1-c)(1-\beta)(1-\eta)}{\delta\beta + (1-c)(1-\beta)(1-\eta)}\frac{1+c}{2} = c \tag{A.47}$$

which can be solved for δ:

$$\delta = (1-\eta)\frac{1-\beta}{\beta}\frac{(1-c)^2}{2c-1} \tag{A.48}$$

The RHS is continuous and strictly monotonically decreasing in c for $c \in [\frac{1}{2}, 1]$. Thus, (A.48) implicitly defines a continuous and strictly monotonically decreasing function $c^A(\delta)$, which attains its maximum at $\delta = 0$. Substituting $\delta = 0$ into (A.48) and solving for c yields $c^A(0) = 1$. As $c^A(\delta)$ is decreasing, we have $c^A(1) < 1$. Figure A.14 illustrates $c^A(\delta)$. The second fact implies

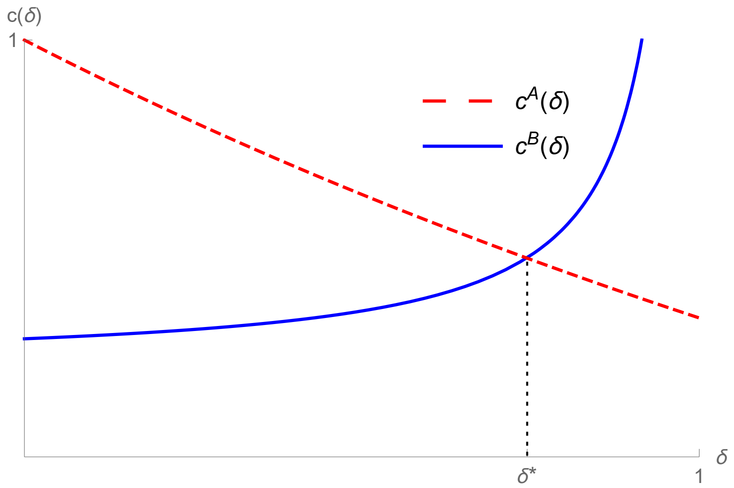


Figure A.14: $c^A(\delta)$ and $c^B(\delta)$ intersect exactly once in $[0,1]$

$$\int_{c}^{1} \eta \frac{1-\beta}{\beta} \frac{m-c}{c-(1-\nu)\bar{q}-\nu m} dm = 1-\delta \tag{A.49}$$

It mirrors the equilibrium condition in the baseline model (see eq. (A.15) 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

$$\int_{c}^{1} \frac{m-c}{c-(1-\nu)\bar{q}-\nu m} dm = \frac{1-\delta}{\eta} \frac{\beta}{1-\beta} \tag{A.50}$$

The RHS of (A.50) is decreasing in δ and, as shown in the proof of Proposition 1, the LHS is decreasing in c. Thus, (A.50) 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 A.14 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 7: Suppose $\nu < \frac{1-\sqrt{\beta}}{1-\sqrt{\beta}}$. This is equivalent to $(1-\nu)\bar{q} + \nu = \underline{c} < c^{JO} = \frac{1}{1+\sqrt{\beta}}$. Recall that (A.48) 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 (A.50) implicitly defines a strictly increasing function $c^B(\delta)$. Let $\langle \eta_n \rangle$ be 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 A.15 and implies that in the limit, as $\eta \to 0$, $\delta^* \to 1$ and $c(\delta^*) \to c^{JO}$.

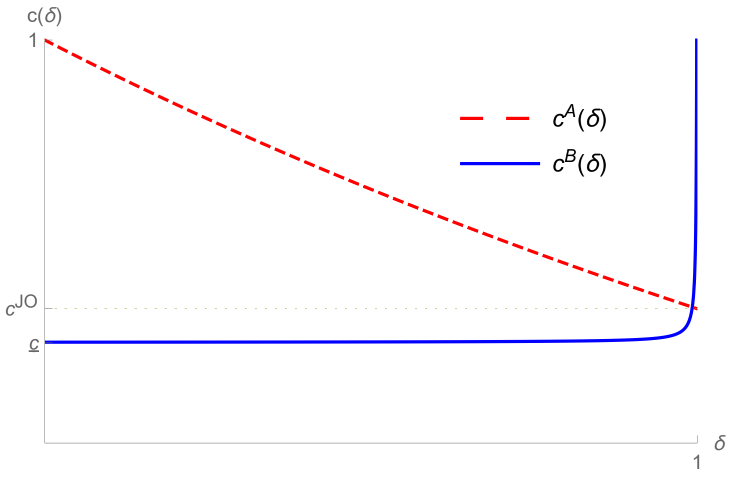


Figure A.15: $c^A(\delta)$ and $c^B(\delta)$ as $\eta \to 0$

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. $\square$

Chapter 2

Gender Differences in Bayesian Updating: Experimental Evidence[1]

2.1 Introduction

Situations of asymmetric information arise in everyday interactions and are relevant in many market contexts. In consumer markets, for instance, producers or sellers typically hold more information about a product than consumers. The emergence of customer reviews ostensibly reduces this asymmetry through the dissemination of information among consumers. However, the presence of fake reviews introduces another dimension of information asymmetry. Not only are consumers uninformed about product quality but, in addition, also about the credibility of the very reviews they consult in order to inform themselves.[2] This results in a two-step problem for consumers. First, they need to form a belief about the trustworthiness of a review. Second, they need to map this belief together with the review's information into a posterior belief about product quality. Similar situations arise in other contexts as well. Consider for example a hiring decision, where the applicant has private information about their ability while the recruiter needs to infer it from a letter of reference. Sometimes these letters are not honestly assessing candidates' abilities but instead try to portray them in a favourable light. Thus, on top of the asymmetry in information about the candidate's type, there is additional uncertainty about the trustworthiness of the source of information which is supposed to resolve the former uncertainty. Financial markets exhibit a similar problem of asymmetric information because credit rating agencies (CRAs) have been found to inflate ratings in order to foster

[1]This chapter is joint work with Dominik Stelzeneder.

[2]Mayzlin et al. (2014) exploit organizational differences between two platforms to show empirically that reviews differ significantly when faking them is made more difficult. He et al. (2021) directly track down recruitment of fake reviewers via Facebook.

business relations (Fulghieri et al., 2014).

Therefore, while reviews, letters of reference, and credit ratings are instruments for resolving information asymmetries, fake reviews, dishonest reference letters, and inflated ratings are detrimental to this objective. Addressing the problem of misinformation directly is a natural approach to the problem. Competition (Board and Lu, 2018; Au and Kawai, 2020, 2021), reputation (Sobel, 1985; Ottaviani and Sørensen, 2006), or regulation (Bourjade and Jullien, 2011; Janssen and Roy, 2022) are all means to reduce misinformation but cannot alleviate the problem completely. Even though the credit rating market is competitive and CRAs have reputation concerns, inflated ratings can still be observed. Despite Amazon's introduction of a verification system and its enormous effort in fighting fake reviews, a substantial share of reviews continues to be fake.[3]

It is thus necessary to understand and mitigate the challenge that the existence of misinformation imposes on consumers. Since the presence of misinformation makes truthful information less informative, even a fully rational decision maker who perfectly understands the problem suffers from it. However, cognitive biases and limitations may exacerbate this problem even more. First, decision makers' beliefs about the trustworthiness of information might be biased or noisy. If consumers believe the share of fake reviews to be higher or lower than it actually is, this might affect the inference they make negatively. Second, misinformation complicates the inference process as it becomes a (more) complex Bayesian updating problem. Mistakes in Bayesian updating might thus have additional negative effects and consumer welfare could decrease even more as a result.

In this paper, we study these negative effects and how to mitigate them. We do this by mimicking informational policies that can be implemented relatively easily and already have been to some extent, for example in the context of online reviews.[4] In particular, we focus on two interventions, each targeting one of the two potential sources of errors described above, namely misspecified beliefs about the trustworthiness of information and mistakes in Bayesian updating. To this end, we conduct an experiment where subjects need to condition their decisions based on potentially misleading information, reflecting a situation of information asymmetry as described above and common in many market contexts.

Our experiment is based on a simple signaling game, where a sender sends an envelope that contains either \$5 or \$0 with equal probability to a receiver. Only the sender observes

[3]According to the monitoring service *Fakespot*, 42% of 720 million assessed Amazon reviews in 2020 were allegedly fake. See https://www.chicagotribune.com/business/ct-biz-amazon-fake-reviews-unreliable-20201020-lfbjdq25azfdpa3iz6hn6zvtwq-story.html, retrieved on June 9, 2022.

[4]Fakespot.com provides an estimate of the share of fake reviews together with a corrected rating for a given product.

the content of the envelope and, before sending it to the receiver, he labels it with one of two messages: "This envelope contains \$5" or "This envelope is empty". Based on this label and knowing that it is not necessarily truthful, the receiver then decides whether to open the envelope or to take a fixed payment instead (her outside option). This decision affects the payoffs of both players. The sender earns a positive payoff only if the receiver opens the envelope and nothing otherwise. The receiver gets her outside option if she decides not to open the envelope and the content of the envelope (either \$5 or \$0) otherwise.

Two features of our experiment allow us to assess the extent to which the receiver deviates from her individually optimal decision. First, we use the Becker et al. (1964) mechanism to elicit subjects' valuation of the envelope.[5] Second, in an additional task we elicit the receiver's *actual valuation* of the envelope, i.e. her valuation if she would have (i) known the exact likelihood that the label was wrong and (ii) been perfectly able to apply Bayes' rule. The difference between a subject's valuation in the first and second tasks defines the *decision error*.

In our experiment, we study how two informational treatments affect receivers' decision errors and implied gains in consumer welfare. In the first treatment $T1$ (*correct beta*), the receivers are informed about the true probability that a label is misleading, which mitigates potential negative effects due to misspecified beliefs. In our second treatment $T2$ (*slider*), we provide receivers with a "slider" that maps any prior belief into a posterior probability and thus reduces possible adverse consequences of incorrect Bayesian updating. Treatment one thus exogenously induces the correct belief and treatment two enables participants to correctly perform Bayesian updating. In a third treatment $T3$ (*full info*), we study the joint effect of these policies by providing them with all information of the first and second treatments.

Based on the previous experimental literature, a clear prediction for the effects of our treatments cannot be made. The effects of the first treatment will depend on the accuracy of the receivers' beliefs about the share of senders that send the honest message. We expect treatment T1 to have a larger effect if the belief accuracy is low and a smaller or even no effect if the beliefs are accurate. The literature finds mixed results for the accuracy of subjects' beliefs.

While Aoyagi et al. (2022) assess the beliefs of subjects in their experiment as accurate, Costa-Gomes and Weizsäcker (2008) and Rey-Biel (2009) find substantial inaccuracies of stated beliefs. The different observations in those papers might, however, be the result of different games that were used in those papers. The former study used a repeated

[5]Before learning the value of her outside option, the receiver had to indicate the smallest outside option that she would be willing to accept for not opening the envelope.

prisoners' dilemma game, whereas in the other two papers behavior in various one-shot 3 x 3 games was analyzed. Since our experiment differs substantially from those three experiments in the literature, no clear prediction for the effects of treatment $T1$ (*correct beta*) can be made. Concerning the effects of the second treatment, the literature offers some guidance as to what can be expected. Bernheim et al., eds (2019) reports a number of studies that document biases in Bayesian updating (e.g. Phillips and Edwards (1966) and Edwards et al. (1968) for two early papers). The provision of a slider that maps any prior belief into a posterior probability helps to prevent these biases. We thus expect treatment $T2$ (*slider*) to decrease subjects' decision errors and thus have a significant negative effect. Regarding gender differences, Jendraszek (2008) studies probabilistic reasoning among future mathematics teachers and finds that men outperform women in tasks that involve conditional probability problems. We thus hypothesize that women benefit more from treatment $T2$ than men.

There is a large and growing literature that studies fallacies in Bayesian updating and the role of beliefs in economic decision making (Bernheim et al., eds, 2019; Nyarko and Schotter, 2002). Our contribution to this literature is fourfold. First, despite extensive research on the role of beliefs and Bayesian updating in economic decision making, we are the first to study the effect of both roles jointly. Our research thus informs policymakers and organizations that aim to improve customers' purchasing decisions. Consider, for instance, the website fakespot.com that checks the reviews for a given listing in an online store like Amazon and estimates the share of fake reviews. Additionally, it computes a corrected rating taking the estimated share of fake reviews into account. It is important to notice that these two pieces of information differ significantly in the cost of provision. Estimating the share of fake reviews typically involves complex language analysis and machine learning techniques and thus requires substantial resources. In contrast, computing a new rating based on an estimated share of fake reviews is a computation that can be done relatively easily and inexpensively. Understanding the extent to which both these measures help consumers is therefore valuable information for implementing such policies.

Second, we contribute to the literature on gender differences in probabilistic reasoning. Many markets are segmented according to gender or exhibit a high concentration of one gender. In such markets, taking into account gender differences is crucial for efficient policy design. And even in markets that are neither segmented nor dominated by one of the genders, understanding gender differences is important for assessing how a certain policy might affect men differently than women. Even though probabilistic reasoning has been studied extensively in the aggregate and gender differences across many other dimensions have gathered attention, gender differences in probabilistic reasoning remain relatively understudied. With the recent surge of information economics and the continuing focus on gender differences not only in economics but also in other social sciences, this comes as somewhat of a surprise. To our knowledge, the only study specifically exploring gender differences in probabilistic reasoning is Jendraszek (2008), who studied to what extent future teachers of mathematics were subject to a range of probabilistic misconceptions. She finds that men outperformed women in tasks involving conditional probability problems. This suggests that Bayesian updating might impair women's decision making to a greater extent. Surprisingly, we find the opposite as male subjects in our experiment benefited significantly more than their female counterparts from a treatment that provides access to a tool that facilitates Bayesian updating.

Third, we add to a growing literature that quantifies individual decision errors and consequent losses in consumer surplus (Carpenter et al., 2021; Harrison et al., 2020b,a). In this literature, individual utility parameters are structurally estimated based on individuals' pretreatment choices. The actual decisions of interest are then compared to the benchmark decisions that maximize the individuals' specific utility functions. In our paper, we *directly* elicit the individually optimal decision of each participant and we compare this decision to the decision of interest. Our conclusions about individuals' decision errors and foregone consumer surplus thus do not depend on a presumed form of utility function. Importantly, this approach allows us to estimate individual decision errors and not only systematic biases on a population level.

Our fourth contribution is to the literature that studies the role of beliefs in games, which can be broadly categorised into two groups. One group of papers estimates beliefs from observed choices (Schneider, 2019; Aguirregabiria, 2021). The other group of papers elicits beliefs and then relates them to subsequent decisions (Nyarko and Schotter, 2002; Costa-Gomes and Weizsäcker, 2008). In our paper, we use a treatment that directly induces the correct belief instead. We then compare subsequent decisions to the decisions of a control group of participants that was not treated. Estimating or eliciting subjects' beliefs is thus not required and a clearer causal link between beliefs and subsequent decisions can be established.

The results of our experiment are surprising in light of the related literature and have important implications for designing informational policies. While we do not find significant treatment effects in the aggregate sample, analyzing men and women separately reveals unexpected differences. Looking at the male subsample, we find that the *full info* treatment $T3$ improved decisions significantly and substantially, reducing the decision error by 34%. This reduction seems to be largely driven by the improvement in Bayesian updating, given that the *slider* treatment $T2$ had an effect of similar size. Treatment $T1$ (*correct beta*) had no significant effect, however, due to the observed carry-over effect of this treatment, we are cautious with interpreting this result. Focusing on the female subsample, we find no significant effects of any of our treatments. This contrasts previous findings about gender differences in Bayesian updating which suggest that the effect of $T2$ should be larger for women than for men. Finally, the effects on expected consumer surplus mirror the ones on decision error. This is to be expected given that the expected consumer surplus is a function of the decision error.

The remainder of the paper is structured as follows. Section 2.2 explains our experimental design. In section 2.3 we introduce the concept of decision error and show how it relates to consumer surplus. In section 2.4, we present the results of the experiment. After first providing some summary statistics, we look into issues that could potentially confound our findings, and finally, we present the treatment effects. Section 2.5 concludes.

2.2 Experimental Design

Our experiment was based on the canonical sender-receiver game. Each sender had private information about the content of an envelope, either \$0 or \$5, and had to send one of two messages to the receiver of the envelope:

A: "This envelope contains \$5."

B: "This envelope is empty."

Subsequently, the receiver decided whether to open it, receiving its content, or sell it for an amount that had been randomly drawn from $[\$0, \$5]$. The sender received \$3 if the envelope was opened and \$0 otherwise. Both, the senders and the receivers were informed about this structure of the game.

We recruited a total of 200 senders and 400 receivers via *Prolific*. The sample consisted of US citizens and was representative with respect to age, ethnicity, and sex. The experiment was programmed in *oTree* (Chen et al., 2016) and took place in January 2022.

2.2.1 Senders

We distributed 200 envelopes - 100 empty, 100 containing \$5 - randomly among 200 senders. Before observing their envelope's content, the senders had to choose between messages A and B for the case that it was empty. In the case that the envelope contained \$5, message A was sent automatically. After they had made a choice, they were informed about their envelope's content and had to fill out a short demographic questionnaire. The envelope, labeled with one of the two messages, was then sent to a randomly selected receiver to be either opened or sold unopened. If the envelope was opened, the sender earned a bonus of \$3 on top of a \$2 participation fee. Otherwise, they received the participation fee but no bonus.

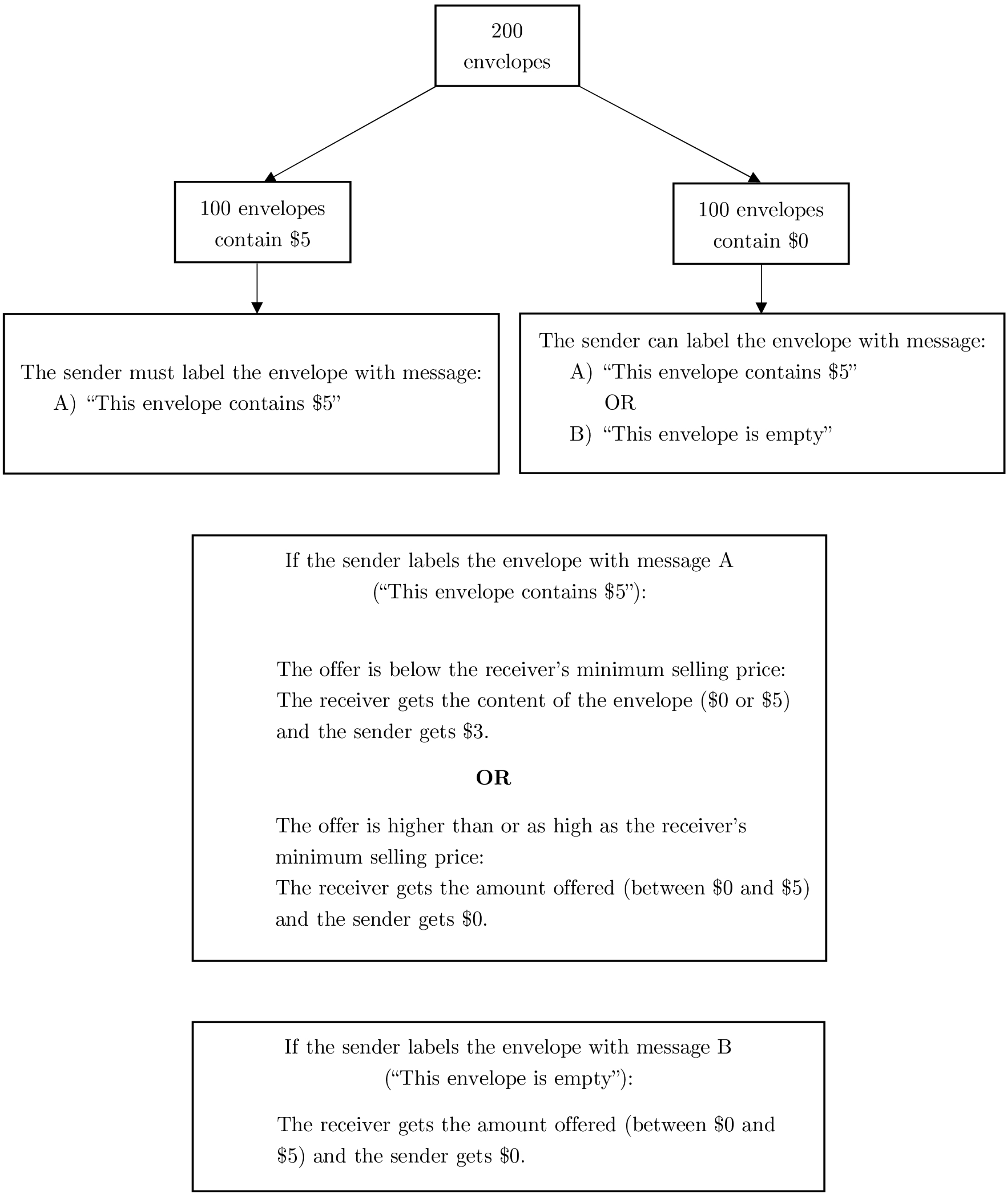


Figure 2.1: Graphical illustration of the sender-receiver game.

2.2.2 Receivers

The receivers' side of the experiment consisted of a total of six tasks. At the end of the experiment, one of the tasks $1 - 4$ was randomly selected for payment. On top of that, Tasks 5 and 6 were both paid out.

After having read the instructions and having seen a graphical illustration of the experiment, participants had to answer six comprehension questions about the experiment. The questions helped to convey the instructions because immediately after submitting their answers, the correct answers to all six questions along with short explanations were provided.

In Task 1, each receiver was randomly assigned one of 200 envelopes. They knew that 100 of them contained \$5, while the remaining ones were empty, and that a sender labeled each of the envelopes according to the rules explained in the section above. Thus, the envelope was labeled either with message A ("This envelope contains \$5.") or with message B ("This envelope is empty."). Before finding out which of the two messages was written on their envelope, they had to name a minimum selling price for the case that it would be message A.[6] Only then the participants were informed about the label of their envelope. If the envelope was indeed labeled with message A, the computer generated a random amount from $[\$0, \$5]$ and offered it in exchange for the envelope. If that offer was at least as high as the stated minimum selling price, the envelope was sold and the receiver was awarded the offered amount. The receiver did not learn about the content of the envelope in this case. Otherwise, if the offer was below the minimum selling price, the envelope was not sold but opened instead. In this case, the receiver got its content, either \$0 or \$5. If, however, the message was B, then the envelope was sold automatically and the receiver was awarded the offered amount. We thus used the Becker-DeGroot-Marschak (BDM, 1964) mechanism to elicit subjects' valuations for an envelope labeled with message A.

In Task 2, each receiver was given a lottery ticket with a winning probability that was based on β, the share of senders who lied. It was calculated to be equal to the posterior probability that an envelope showing message A contained \$5. The outcome of the lottery had two components, a payoff to the receiver and a payoff of \$3 to one sender. If the ticket won, the receiver got \$5, and the sender who got the payoff of \$3 was randomly chosen among those whose envelopes contained \$5. If it lost, the receiver got \$0 and the sender was randomly selected among those whose envelopes contained \$0 and who chose to label them with message A. With the same probability that the envelope in Task 1 showed message B, the lottery ticket was sold automatically. Thus, the lottery was payoff

[6] They did not have to specify a minimum selling price for the case of message B. A letter labeled with message B was sold automatically which is equivalent to a minimum selling price of 0.

equivalent to Task 1. As in Task 1, we elicited subjects' minimum selling price for the lottery ticket via the BDM method. We did not inform participants about the outcome of a lottery if the ticket was sold.

Task 3 consisted of a lottery similar to the one in Task 2, with the only difference that no sender received any payoff. This lottery was thus equivalent to the lottery in Task 2, except for the social externality. As in Tasks 1 and 2, no feedback was given before the end of the experiment.

Task 4 was a cognitive ability test consisting of two parts. The first part was the *Cognitive Reflection Test* introduced by Frederick (2005). It is comprised of three questions, that have intuitive answers, which, however, are incorrect. Only deeper reflection reveals the correct answer. For each correct answer, the participants earned 40 cents. The second part consisted of a series of *Raven's Progressive Matrices* (Raven et al., 1998). The matrices contained either six or nine separate figures that were related to one another in a particular way. In each matrix, one of the nine figures was missing. The participants had to select the correct figure that completed the matrix. Participants had three minutes to complete as many of the 34 matrices as possible and earned 30 cents for every correct solution. Before the timer started, we explained the task with an example.

In Task 5, we used the frequency method (Schlag and Tremewan, 2021) to elicit subjects' beliefs about the share of senders who lied. They received a payoff of \$1 if their answer deviated by less than three from the true β and nothing otherwise.

Task 6 consisted of a series of 20 decision problems. For each of these problems, the participants had to choose between a risky lottery and a safe outcome. The lottery was the same in each decision problem, yielding \$1 with a probability of 75% and \$0 with the remaining probability. The safe outcome was a payoff of 5 cents in the first decision problem and increased by 5 cents for every subsequent problem. Hence, the twentieth problem was a choice between the lottery and a safe payoff of \$1. Of these 20 decisions, one was randomly chosen for payment.

The BDM mechanism used in Tasks $1-3$ is only incentive compatible under the assumption of monotonicity (Azrieli et al., 2018). With Task 6 we can test for strict monotonicity and disregard participants whose preferences do not satisfy this condition from our analyses. To do so, we count the number of decision problems at which a subject switches from choosing the risky lottery to choosing the safe payoff and vice versa. A participant whose utility preferences satisfy the axiom of strict monotonic preferences will either have one or no such switching point. The reason is that after switching from the lottery to the safe payoff once, all subsequent safe payments are larger and thus necessarily preferred given that the subjects has strict monotonic preferences.

Finally, participants had to fill out a short demographic questionnaire before we informed them about all outcomes and their payoffs in each of the chosen tasks. In addition to what they earned in these tasks, they were paid a participation fee of 4$.

2.2.3 Treatments

Our experiment featured a 2×2 between subject design, varying receivers' Task 1. The motivating idea was that for participants to choose optimally in Task 1, two steps were required. First, they needed to have an accurate belief about β, the share of senders who lied. Second, based on their estimate of β, they needed to calculate the posterior probability that an envelope showing message A indeed contained \$5. Our design varied the information provided in Task 1. The baseline treatment $T0$ was exactly as explained above, receivers had no information besides the prior probability that any given envelope contained \$5.

In treatment $T1$ we informed receivers about the share of senders who lied. To this end, we exploited the sequential nature of our experiment, running the sender part first, thus obtaining the realized β before running the receiver part.

In treatment $T2$ we provided participants with a device for the specific purpose of calculating the posterior probability that their envelope was full, respectively empty, given some belief about β. The device in treatment $T2$ took the form of a slider, which they could move to positions from 0% to 100% representing beliefs about β. The formulas for Bayesian updating that the slider applied were as follows.

$$\pi(\$5|A) = \frac{1}{1+\beta} \quad (2.1) \qquad\qquad \pi(\$0|A) = \frac{\beta}{1+\beta} \quad (2.2)$$

The slider applied Equation (2.1) to calculate the posterior probability that the envelope contains \$5, given message A and the participant's belief about β. With Equation (2.2), the slider calculated the posterior probability that the envelope was empty, given the same conditions. Below the slider, participants were informed about the outcomes of the calculations. Hence, they could see the probabilities that the envelope was filled or empty, respectively, conditional on the belief at which they positioned the slider.

Treatment $T3$ combined both pieces of information from the previous treatments. Participants were informed about the correct β *and* they could use a slider mapping prior beliefs into posteriors (see Figure 2.2). Thus, $T3$ eliminated the strategic uncertainty and computational complexity of Task 1 and rendered it equivalent to the lottery in Task 2. Table 2.1 summarizes the treatment design.

Task 1

By clicking on the ‚+' symbols you can access the instructions again. Please take your time and consider your decision carefully.

+ Information on Task 1 (senders)

+ Information on Task 1 (receivers)

+ Graphical Illustration

Additional Informationen:

57% of all senders whose envelopes were empty labelled them with message A ("This envelope contains $5")

Below you can use a calculator that was designed specifically for this task. You can submit an estimate of the share of *senders* who labelled their empty envelopes with message A ("This envelope contains $5"). Based on your estimate, the calculator computes the probability that *your* envelopes contains either $0 or $5. You can submit multiple estimates and use the calculator as often as you wish.

Input:

If 57 % of those *senders* whose envelopes were empty, labelled them with message A ("This envelope contains $5") and you receive an envelope that is labelled with message A, then:

Result:

- Your envelope contains $5 with a probability of 64 %.
- Your envelope contains $0 with a probability of 36 %.

Please answer the following question:

If your envelope is labeled with the message A ("This envelope contains $5"), for how much money would you be willing to sell it to us?

Your answer: $1.8

If we offer you between $0 and $1.79 you will open the envelope.

If we offer you $1.8 or more you will sell the envelope.

Next

Figure 2.2: Screenshot of Task 1 with information on both the share of senders who lied and the slider.

Table 2.1: Treatments

		correct β	
		no	yes
slider	no	*baseline*	*correct beta*
	yes	*slider*	*full info*

Notes. Table shows experimental design.

2.3 Decision Error and Welfare

Our main research question is whether and how our treatments affect men's and women's decision error in Task 1. Judging the error of a decision requires comparing the actual decision of a participant with the "optimal" decision that he or she "should have made".

Carpenter et al. (2021), Harrison et al. (2020b), and Harrison et al. (2020a) are the first that quantify decision errors and the consequent losses in expected consumer surplus. In their experiments, they use participants' pretreatment choices to structurally estimate the parameters of a presumed utility function for each individual. The actual posttreatment decisions of interest are then compared to the benchmark decisions that maximize the individuals' specific utility functions.

In our paper we use a slightly modified approach that does not require structural estimations or assumptions about the specific form of a utility function but instead provides the individually optimal decision directly. In our study, a participant's decision in Task 2 defines her individually optimal decision for Task 1. Recall that Task 2 was identical to Task 1 in terms of outcome. The only difference was that in Task 2 all relevant information for the decision problem was available, which was not necessarily the case in Task 1. Under the assumption of consistency, the decision in Task 2 informs us about the actual valuation of the envelope in Task 1. This is also what a participant *should have chosen* as her minimum selling price in Task 1 if the additional information would have been available to her.

The validity of this experimental approach depends on the assumption that effects of our treatments in Task 1 do not carry over to decisions in Task 2. Such carry-over effects are a well-known issue in experimental within-subject designs (Greenwald, 1992; Charness et al., 2012) and we will test for them in section 2.4.2.

The main advantage of our approach compared to other approaches in the literature is that our results are independent of the specific utility functions of our participants. We can assess the size of the decision error by directly comparing the decisions made in Tasks 1 and 2. If both decisions coincide, then the decision in Task 1 was individually optimal and the error was *zero*. If, however, the decisions do not coincide, then the decision in

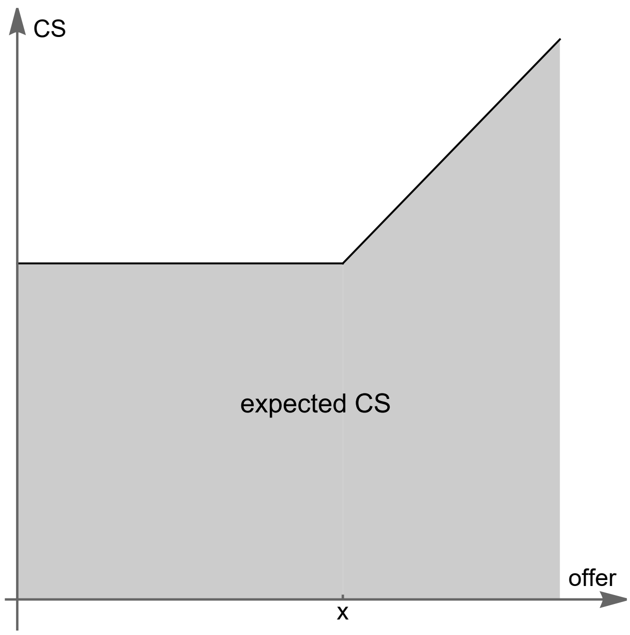


Figure 2.3: Expected Consumer Surplus in the case where the MSP is equal to the valuation.

Task 1 contained an error equal to the extent of the deviation.

The economic consequences of such deviations can be illustrated by a simple welfare analysis. We will see that the size of a decision error has direct implications for a participant's expected consumer surplus (CS).

The calculation of a participant's expected CS is straightforward. Let us first assume that the decision error is 0. This means that a subject's valuation of the envelope, x, and her stated minimum selling price (MSP), y, are both equal, i.e. $y = x$. In this case, the expected CS can be calculated as follows: She obtains a surplus of x whenever the randomly drawn offer in exchange for the envelope is below x and the money offered otherwise. Recall that the offer is a randomly drawn amount from the interval $[0, 5]$. Let us denote the offer by z such that $F(z) = \frac{z}{5}$. Then,

$$\begin{aligned} CS_{no\_error} &= \int_0^x x dz + \int_x^5 z dF(z) = F(x)x + \int_x^5 \frac{z}{5} dz = \frac{x^2}{5} + \frac{z^2}{10}\Big|_x^5 \\ &= \frac{2x^2}{10} + \frac{5^2}{10} - \frac{x^2}{10} = \frac{x^2 + 5^2}{10} \end{aligned}$$

Given that a subject's valuation of the envelope lies in $[0, 5]$, CS_{no_error} takes its minimum at $x = 0$ where it is 2.5 and its maximum at $x = 5$ where it is 5. Figure 2.3 illustrates the CS in the case where MSP and valuation coincide, i.e. without decision error.

Let us now turn to the case in which subjects make decision errors by submitting MSPs that are not equal to their valuations of the envelope. As before, x denotes a subject's valuation for the envelope and y their MSP, only now $y \neq x$. Then, the expected CS in Task 1 can be computed as:

$$CS_{with\_error} = \int_0^y x dz + \int_y^5 z dF(z) = F(y)x + \int_y^5 \frac{z}{5} dz = \frac{xy}{5} + \frac{z^2}{10}\bigg|_y^5$$
$$= \frac{2xy}{10} + \frac{5^2}{10} - \frac{y^2}{10} = \frac{x^2-(x-y)^2}{10} + \frac{5^2}{10} = CS_{no\_error} - \frac{(x-y)^2}{10}$$

And the foregone consumer surplus is simply:

$$CS_{foregone} = \frac{(x-y)^2}{10} \tag{2.3}$$

Thus, the loss in consumer surplus is proportional to the decision error squared. Graphically, this can be illustrated as in Figure 2.4. The left panel depicts the case where the MSP is below the actual valuation of the envelope. In this case, whenever the offer is between the MSP and the valuation, the participant receives y although she would have valued opening the letter with $x > y$. Mathematically, the loss in expected CS is the integral of the difference over all such offers, weighted by their probability densities. This corresponds to the shaded area. The case where the MSP is above the valuation is depicted in the right panel of Figure 2.4. In both cases, it is thus the difference between valuation and MSP, i.e. the *decision error*, that determines the loss in expected CS.

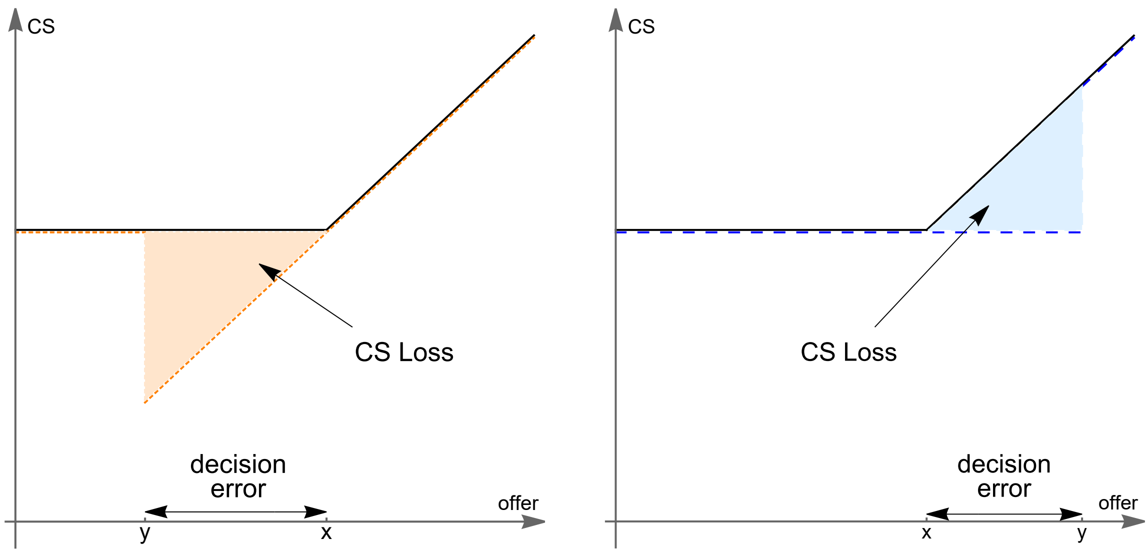


Figure 2.4: The shaded areas illustrate the loss in CS due to decision error. The left(right) panel depicts the case where the MSP is lower(higher) than the valuation.

2.4 Results

We will present the results of this study in the following sequence. In section 2.4.1 we summarize the decisions of the senders and present the implications for the receivers in the main part of the experiment. Section 2.4.2 is concerned with issues that could arise if subjects violate the axiom of monotonic preferences or if treatment effects in Task 1 carry over to Task 2. Finally, section 2.4.3 presents the effects of our treatments on participants' decision errors and consumer welfare.

2.4.1 Senders' decisions

This section provides the results for the senders' decisions. We will use these results to compute the winning chances of the lottery tickets for the receivers in Task 2.

In the first part of the experiment, the senders had to choose a label for their envelope. 57% of all senders who received an empty envelope labeled it with message A ("This envelope contains \$5."). Envelopes that contained \$5 were automatically labeled with message A.

What does this mean for the receivers? First, with a probability of 0.215 the envelope in Task 1 was sold automatically because it showed message B ("This envelope is empty."). Second, the share of senders that labeled an empty envelope with message A determines the probability that an envelope with message A contained \$5. According to Equation (2.1), an envelope that showed message A contained \$5 with a probability of 0.64.

Our experimental design requires outcome equivalence between Task 1 and the lottery in Task 2. We obtain this outcome equivalence by assigning a winning chance of 0.64 to the lottery ticket. In addition, the probability that the lottery ticket is sold automatically was set to 0.215. This is identical to the chance that the envelope in Task 1 was labeled with message B and thus sold automatically.

2.4.2 Monotonicity and Carry-Over Effects

This section reports test results for two potential issues that could arise in our experiment and lead to biased estimates of treatment effects. The first issue concerns violations of the monotonic preference assumption while the second issue relates to carry-over effects. Therefore, we report our results on these issues before proceeding with presenting the main results of our experiment in the next section.

We first report the results of our test for violations of monotonic preferences. If subjects violate the monotonic preference assumption, we have to drop them from the main analysis. The reason is that the BDM mechanism that we use in Tasks $1-3$ is only incentive compatible under this assumption (Azrieli et al., 2018). A subject with strictly monotonic preferences can have at most one switching point from the lottery to the safe outcome. Thus, if subjects chose the safe outcome in a decision problem and the lottery in any of the subsequent decision problems (with strictly higher safe outcome), they violated

Table 2.2: Effects of Treatment in Task 1 on Decisions in Task 2

Task 2 Decision	Mean	N	Treatment 0	Treatment 1	Treatment 2
Treatment 0	3.26	93			
Treatment 1	2.97	95	$p = 0.071$		
Treatment 2	3.10	91	$p = 0.349$	$p = 0.277$	
Treatment 3	3.22	86	$p = 0.902$	$p = 0.071$	$p = 0.384$
Whole sample	3.13	365			joint F-test: $p = 0.212$

Notes. Table shows mean answers in Task 2 by treatment in Task 1. p-values are based on t-tests. N denotes sample size.

the strict monotonicity assumption. In our experiment, 35 participants made choices that were inconsistent with this assumption and were thus excluded from the analysis.[7]

Next, we turn to reporting our findings regarding carry-over effects. Carry-over effects occur when a treatment affects the decision of a participant in an unrelated subsequent task (Greenwald, 1992; Charness et al., 2012). In our experiment, this means in particular that the unrelated treatment in Task 1 could have affected decisions in Task 2. In this case, our judgment about a participant's decision error could be distorted and we would, therefore, not be able to perform unbiased analyses of the concerned treatment.

We now report two tests for carry-over effects. First, we regress the decisions in Task 2 on the categorical treatment variable. The treatment should, since it is unrelated to Task 2, not affect decisions in Task 2. We thus expect not to find significant differences in means across the four treatments. The results of this first test are shown in Table 2.2. We see that treatment $T1$ (*correct beta*) had a marginally significant effect on decisions in Task 2 (-0.34 compared to treatment $T0$ and -0.32 compared to treatment $T3$, both $p = 0.071$). Decisions did not significantly differ between treatment $T0$ (baseline), treatment $T2$ (slider), and treatment $T3$ (full info). Although the joint F-test does not reject the *null* ($p = 0.212$), it seems possible that treatment $T1$ affected decisions in Task 2.[8] Such a potential carry-over effect from treatment $T1$ to answers in Task 2 might distort our measure of the decision error. Therefore, treatment effect estimates for treatment $T1$ have to be interpreted with caution.

Second, we test whether our treatments affected the share of participants who made two identical decisions in Tasks 1 and 2 due to a "taste for consistency" citepFalk2012. The concern could be that due to treatment in Task 1, subjects might have adjusted their decision because they recognized that Tasks 1 and 2 were identical in terms of outcomes. Table 2.3 reports the share of decisions that coincide, respectively differ, by treatment

[7]Note that our results would not change qualitatively if they were included in the analyses.

[8]Results of our balance checks do not suggest that a randomization issue caused the observed differences.

Table 2.3: Share of participants whose decisions in Tasks 1 and 2 coincide

Tasks 1 and 2	Treatment 0	Treatment 1	Treatment 2	Treatment 3
Decisions coincide	32%	42%	34%	36%
Decisions differ	68%	58%	66%	64%
N	93	95	91	86

$p = 0.532$, Fisher's exact test

status. We see that these shares are not significantly affected by the treatments. Neither the joint test ($p = 0.532$, Fisher's exact test) nor *Wald* tests that compare all treatments pairwise are significant at conventional significance levels. We thus do not find evidence for a carry-over effect of the treatments on the share of coinciding decisions in Tasks 1 and 2.

What do these results mean for our main analyses? First, we will disregard 35 participants from the analyses because of violations of the monotonicity assumption. Second, we will interpret results for treatment $T1$ with great caution because the effects of this treatment might have carried over to decisions in Task 2.

2.4.3 Decision Error and Welfare

In this section we present our main results. Table 2.4 shows how each treatment affected the participants' decision errors. Recall that the decision error is defined as the absolute difference between a subject's decisions in Tasks 1 and 2. The data are thus censored from below at 0. The estimates are based on a simple Tobit model with the decision error as the dependent variable and a categorical treatment indicator as the independent variable.

When the decisions of men and women are analyzed jointly, we do not see significant effects of our treatments. Even though all three treatments reduced the decision error, the point estimates are not significant ($T1$: -0.177, $p = 0.108$; $T2$: -0.121, $p = 0.281$; $T3$: -0.162, $p = 0.149$).

Analyses of the treatment effects on men and women separately, however, reveal surprising findings. First, women's decision error does not seem to decrease due to our treatments. Women in treatment $T3$ who received full information, i.e. information about the exact share of senders who lied as well as a tool to map this share into the posterior (thus reducing the problem to a simple lottery), did not make significantly better decisions than women who did not receive any additional information (-0.043, $p = 0.781$). The point estimates for the effects of treatments $T1$ (*correct beta*) and $T2$ (*slider*) are insignificant as well ($T1$: -0.140, $p = 0.334$; $T2$: 0.092, $p = 0.587$).

Second, men could improve their decisions in the full information treatment $T3$ (-0.287,

$p = 0.075$). The average gap between decisions in Tasks 1 and 2 decreased by 34% relative to the gap in the baseline treatment $T0$. It seems that this improvement was largely driven by the effect of the slider, which, on its own, had an effect of almost identical size on the decision error (treatment $T2$: -0.298, $p = 0.055$). Treatment $T1$ did not reduce men's decision error significantly.[9]

When we compare the treatment effects on men and women, we see that the slider has a significantly different effect on the two genders. Men benefit more from this treatment than women (0.390, $p = 0.079$, difference-in-differences). The effect of treatment $T3$ (full info) is not significantly different for men and women (0.239, $p = 0.298$, difference-in-differences).

The implications for expected consumer welfare are shown in Table 2.5. In the baseline treatment $T0$, participants left an average consumer surplus of \$0.142 on the table. The decreases in foregone consumer surplus due to treatments $T1$ to $T3$ are not significant in the pooled sample ($T1$: \$$-0.041$, $p = 0.131$; $T2$: \$$-0.040$, $p = 0.133$; $T3$: \$$-0.042$, $p = 0.124$).

In the separate analyses for both genders, we see that men were able to significantly decrease foregone consumer surplus by \$0.095 ($p = 0.025$) in treatment $T2$ and by \$0.091 ($p = 0.038$) in treatment $T3$. These are each reductions of almost half of foregone consumer surplus compared to men's baseline level in treatment $T0$.[10] The reduction of \$0.075 ($p = 0.098$) in treatment $T1$ has to be interpreted with a grain of salt due to the observed carry-over effects. Women, in contrast, do not seem to benefit from the treatments, all three estimates are insignificant ($T1$: \$$-0.012$, $p = 0.697$; $T2$: \$0.016, $p = 0.638$; $T3$: \$0.004, $p = 0.894$). The availability of the slider in treatment $T2$ helped men to reduce foregone consumer surplus by \$0.109 more than women ($p = 0.049$). We do not see a significant gender difference in the effect of treatments $T1$ (\$0.052, $p = 0.405$) or $T3$ (\$0.090, $p = 0.118$). The results on consumer surplus thus mirror the ones on decision errors and show that the mistakes that subjects made directly translated into losses in consumer surplus.

[9] Remember that treatment $T1$ is potentially compromised due to carry-over effects.

[10] Note that men's foregone consumer surplus in the baseline treatment $T0$ is slightly larger in our sample than women's foregone consumer surplus. The difference is, however, not statistically significant ($p = 0.165$).

Table 2.4: Unconditional Marginal Effects on Decision Error

	All	Men	Women	Gender difference
Decision error	(1)	(2)	(3)	(4)
Treatment 1 (*beta*)	-0.177	-0.221	-0.140	0.065
	(0.110)	(0.165)	(0.145)	(0.169)
Treatment 2 (*slider*)	-0.121	-0.298*	0.092	-0.390*
	(0.113)	(0.155)	(0.168)	(0.223)
Treatment 3 (*full info*)	-0.162	-0.287*	-0.043	0.239
	(0.112)	(0.161)	(0.156)	(0.229)
Baseline	0.766***	0.852***	0.689***	-0.153
Treatment 0	(0.095)	(0.163)	(0.097)	(0.178)
N	365	187	178	

Notes. This table reports unconditional average marginal treatment effects on decision error. Estimates are based on a Tobit regression model, data are censored from below at 0. Dependent variable is *decision error*, which is the absolute difference between the answers of a participant in Task 1 and Task 2. Standard errors in parentheses. Column *Gender difference* reports difference-in-differences estimates for gender differences in treatment effects. *Baseline* reports mean decision errors in treatment $T0$ with bootstrapped standard errors in columns (1) to (3).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.5: Unconditional Marginal Effects on Foregone Consumer Surplus

Decision error	All (1)	Men (2)	Women (3)	Gender difference (4)
Treatment 1 (*beta*)	-0.041	-0.075*	-0.012	0.052
	(0.027)	(0.045)	(0.030)	(0.062)
Treatment 2 (*slider*)	-0.040	-0.095**	0.016	0.109**
	(0.027)	(0.042)	(0.034)	(0.055)
Treatment 3 (*full info*)	-0.042	-0.091**	0.004	0.090
	(0.027)	(0.044)	(0.032)	(0.058)
Baseline	0.142***	0.194***	0.095***	-0.069
Treatment 0	(0.034)	(0.068)	(0.019)	(0.050)
N	365	187	178	

Notes. This table reports unconditional average marginal treatment effects on expected foregone consumer surplus. Estimates are based on a Tobit regression model, data are censored from below at 0. Dependent variable is *foregone consumer surplus*, as calculated in equation 2.3. Standard errors in parentheses. Column *Gender difference* reports difference-in-differences estimates for gender differences in treatment effects. *Baseline* reports mean foregone consumer surplus in treatment $T0$ with bootstrapped standard errors in columns (1) to (3).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.5 Conclusion

In this paper, we study how different informational treatments affect the decision error of men and women in a simple signaling game. The first treatment is designed to alleviate potential negative effects of misspecified beliefs. The second treatment reduces the decision error due to incorrect Bayesian updating. Finally, the third treatment combines all elements of the former two.

We find strong evidence for gender differences in the effects of our treatments. While men's decision quality is significantly improved by a "slider" that maps a prior belief into a posterior probability, we do not find this treatment to affect women's decision errors. Additional information about the actual share of senders who do not send the true message did not lead to a further decrease in decision error for neither gender. Welfare analyses show a substantial decrease in men's foregone consumer surplus due to the slider. Their foregone consumer surplus of $0.194 was almost halved by our treatment.

The significance of our results is two-fold. On the one hand, we make an important contribution to the literature about gender differences in probabilistic reasoning by showing that the inability to correctly calculate conditional probabilities affected men more strongly than women. This is in contrast to previous findings in the literature, which point in the opposite direction. Overall, empirical evidence in this domain is scarce and, as we show, far from conclusive. On the other hand, the findings in this paper make an important contribution to the ongoing discussion about regulating online platforms. In light of the pervasiveness of fake reviews, awareness policies have been discussed as a concrete measure for alleviating the negative effects they have on consumers.[11] Estimates about the share of fake reviews as provided by websites like Fakespot.com are one natural way of providing this information. The results in this paper highlight the importance of complementing this information with a corrected rating. In fact, we find that a mere tool that helps calculate corrected ratings given individuals' estimates of the share of fake reviews has the potential to be as effective as both pieces of information. Policymakers need to bare this in mind when designing awareness policies.

Another implication of our findings is that such policies do not have the same effect on women as they have on men. In particular, this study suggests that while such interventions might be an effective consumer protection policy in male-dominated markets, they may not be very effective in female-dominated markets.

[11] On a European level such measures have been brought forward by the European Commissioner for justice, who insists that businesses provide visible information about the reliability of reviews.

2.A Instructions for Senders

The sender instructions below do not correspond exactly to what is described in section 2.1. They include an additional task that is unrelated to the experiment itself and we want to briefly explain why we included it.[12] In order to have access to a representative sample *Prolific* requires a minimum of 300 participants per study. Given that our experiment consisted of two studies (the sender part and the receiver part), we had to include 300 subjects in the sender part. To keep the experiment simple and the instructions clear, we matched every sender with two receivers. One for whom Task 1 or 2 and thus the envelope was relevant, and one for whom Task 3 or 4 was relevant. That way, we ensured that each sender sent his envelope to *one* receiver whose decision affected the sender's payoff. It was therefore important to have twice as many receivers as senders.

For budgetary reasons, the only way to conduct our study with a representative sender sample and to maximise the number of receivers while still maintaining a one-to-two matching, was to recruit 300 sender of which we matched only 200. Of those 300 senders, 100 got assigned an empty envelope, 100 one that contained $5, and 100 were assigned to the additional task.

[12]The task consisted of the same *Cognitive Reflection Test* that was part of receivers' Task 4.

Welcome!

In today's experiment you will be asked to complete three tasks. This will take no longer than 10 minutes and your decisions in these tasks will remain anonymous. At the end of the experiment, you will receive a participation fee of $2 for having completed the three tasks. In addition, we will randomly select one of the three tasks to be relevant. Before each task we will describe in detail how your payment is determined within that task.

Your total earnings from the experiment are the sum of your participation fee and your bonus payment from the randomly selected task.

Consent form and data protection information

The University of Vienna, Universitaetsring 1, 1010 Vienna, Austria is conducting a scientific study using Prolific today.

Your responses will be recorded on our server. The data generated in the study will be separated from the data in the Prolific system after the payment has been completed. The final dataset will not allow any inference to be drawn about the individual participants or the responses of those participants that can be linked to their identity in any way. Correspondingly, the analysis and presentation of all results of this experiment will be anonymized. The anonymous research data will be archived and will possibly be made available to other scientists for further use.

We want to ask for your consent to participate in the experiment and to the implied processing of data. You can revoke your consent at any time, for example, by sending an e-mail to dominik.stelzeneder@univie.ac.at. Your consent is at the same time the legal basis for the processing of data.

For any general information about the processing of data and data protection, especially about your rights to disclosure, correction, deletion, limitations of processing of data, objection, data transmission and complaints to a supervisory body, please contact the data protection officer of University of Vienna: verarbeitungsverzeichnis@univie.ac.at. You also have the right of complaint to the Austrian Data Protection Agency (dsb@dsb.gv.at).

Participation in today's study is entirely voluntary. You have the possibility to exit the study at any time.

You cannot participate in this experiment without consenting.

I have read the data protection information and I consent to participate in the experiment and to the described data processing procedures:

- ○ Yes
- ○ No

Tasks 1 and 2

For this experiment we prepared 200 envelopes in the following way: We put \$5 in 100 envelopes. The remaining 100 envelopes remain empty, they thus contain \$0.

There are two distinct groups of participants in this experiment, *senders* and *receivers*. <u>You are one of the *senders*</u> (and this will not change throughout the experiment). Each *sender* receives one randomly selected envelope and they are asked to label it with one of two messages:

A) “This envelope contains \$5”
B) “This envelope is empty “

The rules are the following:
If the envelope <u>contains \$5,</u> you can <u>only choose</u> message A. If, however, the envelope <u>is empty,</u> you can <u>choose freely</u> one of the two messages.

Subsequently, the envelope will be sent to a randomly selected *receiver*. For the *receiver* the experiment will proceed as follows. Each *receiver* will have to name the lowest offer they are willing to accept in exchange for the envelope – the *minimum selling price*. The computer will then generate a random amount between \$0 and \$5 and offer it to the *receiver*.

If the envelope is labelled with message A (“This envelope contains \$5”), the *receiver* either opens or sells the envelope:

- If the computer’s offer is below the *minimum selling price*, the *receiver* opens the envelope. In this case the *receiver* receives whatever amount the envelope contains and you get \$3.
- If the computer’s offer is higher than or as high as the *minimum selling price*, the *receiver* sells the unopened envelope. In this case the *receiver* receives the offered amount and you get \$0.

If the envelope is labelled with message B (“This envelope is empty”), the *receiver* sells the envelope irrespective of the submitted *minimum selling price*.

- In this case the *receiver* receives the offered amount and you get \$0.

These instructions are provided to all participants of the experiment.

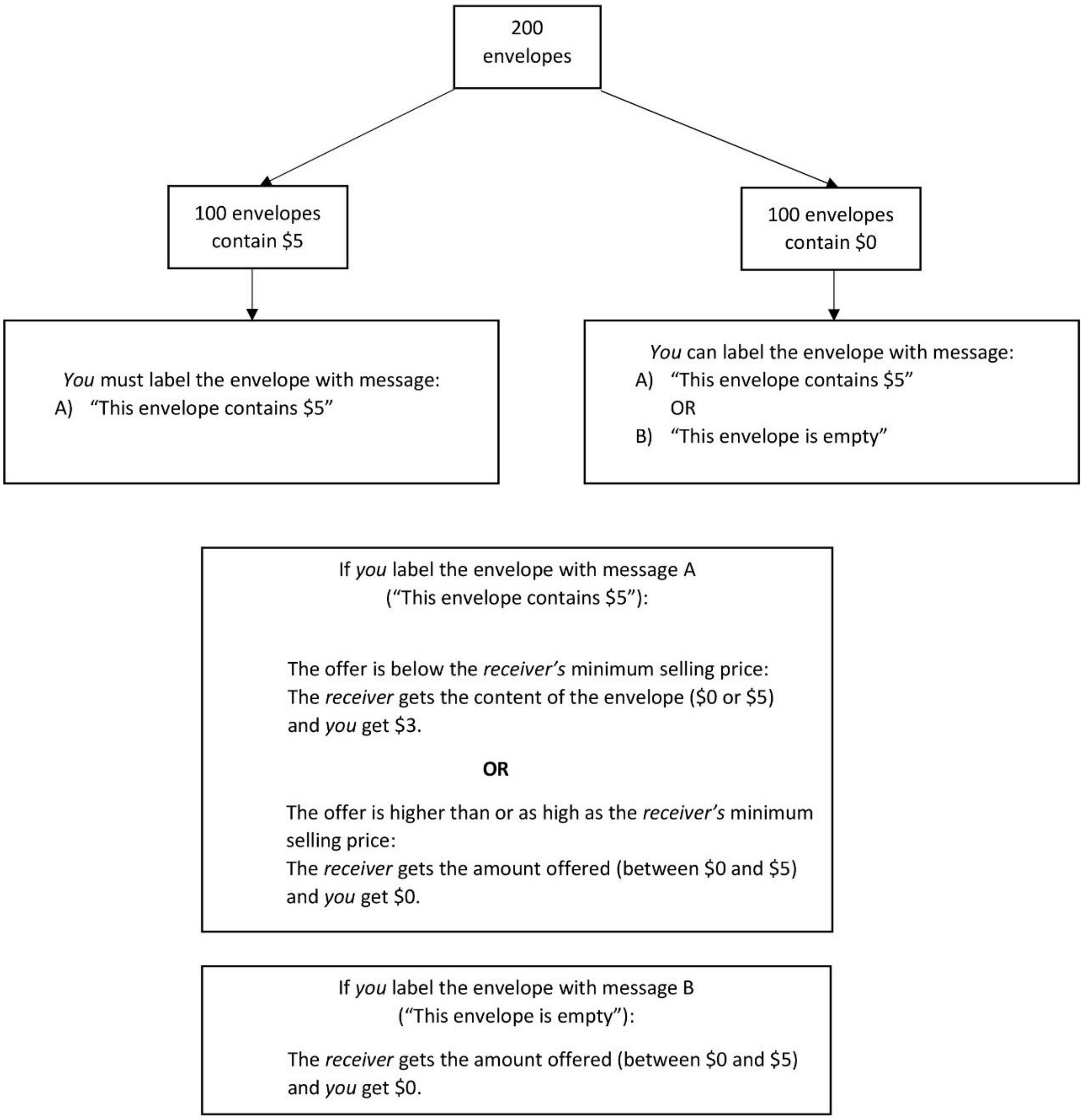
Your task will be to label your envelope.

If your letter contains \$5, it will automatically be labelled with message A (“This envelope contains \$5”).

If it is empty, you will be asked to answer the following question:

Which message do you choose to label your envelope with?

Tasks 1 and 2



In Task 1 on the next page you will answer the following question:

Which message do you choose to label your envelope with?

Task 1 - Decision

By clicking on the ‚+‘ symbols you can access the instructions again. Please take your time and consider your decision carefully.

+ Information on Task 1

+ Graphical Illustration

Your envelope is empty.

Please answer the following question:

Which message do you choose to label your envelope with?

Message A: “This envelope contains $5”
Message B: “This envelope is empty”

Please click ‘Next’ after you have selected a message.

Task 2 - Decision

Your envelope contains 5 $.

Your envelope will therefore automatically be labelled with message A ("This envelope contains $5").
Please click 'Next' to proceed.

Task 3

Please answer the following questions. You will receive $0.40 for each correct answer if Task 3 is randomly selected to be relevant.

Question 1:
A bat and a ball together cost $1.10. The bat is $1 more expensive than the ball. How much does the ball cost?

Answer (in cents):

Question 2:
5 machines take 5 minutes to produce 5 items. How long do 100 machines take to produce 100 items?

Answer (in minutes):

Question 3:
A lake is partially covered with lily pads. The area that is covered by lily pads doubles every day. If it takes 48 days for the lily pads to cover the whole lake, how long does it take until half of the lake is covered?

Answer (in days):

Final Questions

What is your gender?

- Woman
- Man
- Non-binary
- Prefer not to say

How old are you?

What is your highest completed level of education?

Thank you for your participation!

You earn $2 for partaking in today's experiment. Task 2 was randomly selected to be relevant, you thus received an envelope that contained $5.

Hence, you automatically labelled it with message A ("This envelope contains $5"). Therefore, your additional earnings depend on the *receiver*'s decision. If the receiver opens the envelope, you will get a bonus of $3. Otherwise, you will not get a bonus. Hence, you will receive either $2 or $5 in total. We will send you your participation fee within two days. As soon as the *receiver* of your envelope has made a decision and your bonus for Task 2 has been determined, we will send you your bonus. This should not take longer than a week.

Feedback

Lastly, we kindly ask for your feedback on this study. Was anything unclear? What do you think could be improved?

2.B Instructions for Receivers

Note also that in the following receiver instructions, the depicted practice exercise (p.88 – 89) does not show the actual Raven matrix that we used in the experiment.

Welcome!

In today's experiment you will solve 4 Tasks. Each task will take less than 15 minutes and your decisions will remain anonymous. At the end of the experiment, you will receive a participation fee of $ 4 for having completed the four tasks. In addition, we will randomly select one of the tasks and pay you based on your decisions or your performance in that task. In addition, there will be two additional tasks, that will contribute to your earnings. Before each task we describe in detail how your payment is determined within that task. At the end of the experiment, we inform you about which task was selected, and we break down your total earnings. They will represent the sum of your participation fee, your earnings in the randomly selected task, and your earnings in the additional tasks.

Consent form and data protection information

The University of Vienna, Universitaetsring 1, 1010 Vienna, Austria is conducting a scientific study using Prolific today.

Your responses will be recorded on our server. The data generated in the study will be separated from the data in the Prolific system after the payment has been completed. The final dataset will not allow any inference to be drawn about the individual participants or the responses of those participants that can be linked to their identity in any way. Correspondingly, the analysis and presentation of all results of this experiment will be anonymized. The anonymous research data will be archived and will possibly be made available to other scientists for further use.

We want to ask for your consent to participate in the experiment and to the implied processing of data. You can revoke your consent at any time, for example, by sending an e-mail to dominik.stelzeneder@univie.ac.at. Your consent is at the same time the legal basis for the processing of data.

For any general information about the processing of data and data protection, especially about your rights to disclosure, correction, deletion, limitations of processing of data, objection, data transmission and complaints to a supervisory body, please contact the data protection officer of University of Vienna: verarbeitungsverzeichnis@univie.ac.at. You also have the right of complaint to the Austrian Data Protection Agency (dsb@dsb.gv.at).

Participation in today's study is entirely voluntary. You have the possibility to exit the study at any time.

You cannot participate in this experiment without consenting.

I have read the data protection information and I consent to participate in the experiment and to the described data processing procedures:

- ○ Yes
- ○ No

Task 1

For this experiment we prepared 200 envelopes in the following way: We put $5 in 100 envelopes. The remaining 100 envelopes remain empty, they thus contain $0.

There are two distinct groups of participants in this experiment, *senders* and *receivers*. Each *sender* receives one randomly selected envelope, and they are asked to label it with one of two messages:

A) “This envelope contains $5”
B) “This envelope is empty”

The rules are the following: if the envelope contains $ 5, the *sender* can only choose message A. If, however, the envelope is empty, the *sender* can choose freely one of the two messages.

Subsequently, the envelope will be sent to a randomly selected *receiver*.

You are one of the *receivers*. You will have to name the lowest offer that you are willing to accept in exchange for the envelope – the *minimum selling price*. The computer will then generate a random amount between $0 and $5 and offer it to you.

If the envelope is labelled with message A (“This envelope contains $5”), you either open or sell the envelope:

- If the computer’s offer is below the *minimum selling price*, you open the envelope. In this case you receive whatever amount the envelope contains, and the *sender* gets $3.
- If the computer’s offer is higher than or as high as the *minimum selling price*, you sell the unopened envelope. In this case you receive the offered amount, and the *sender* gets $0.

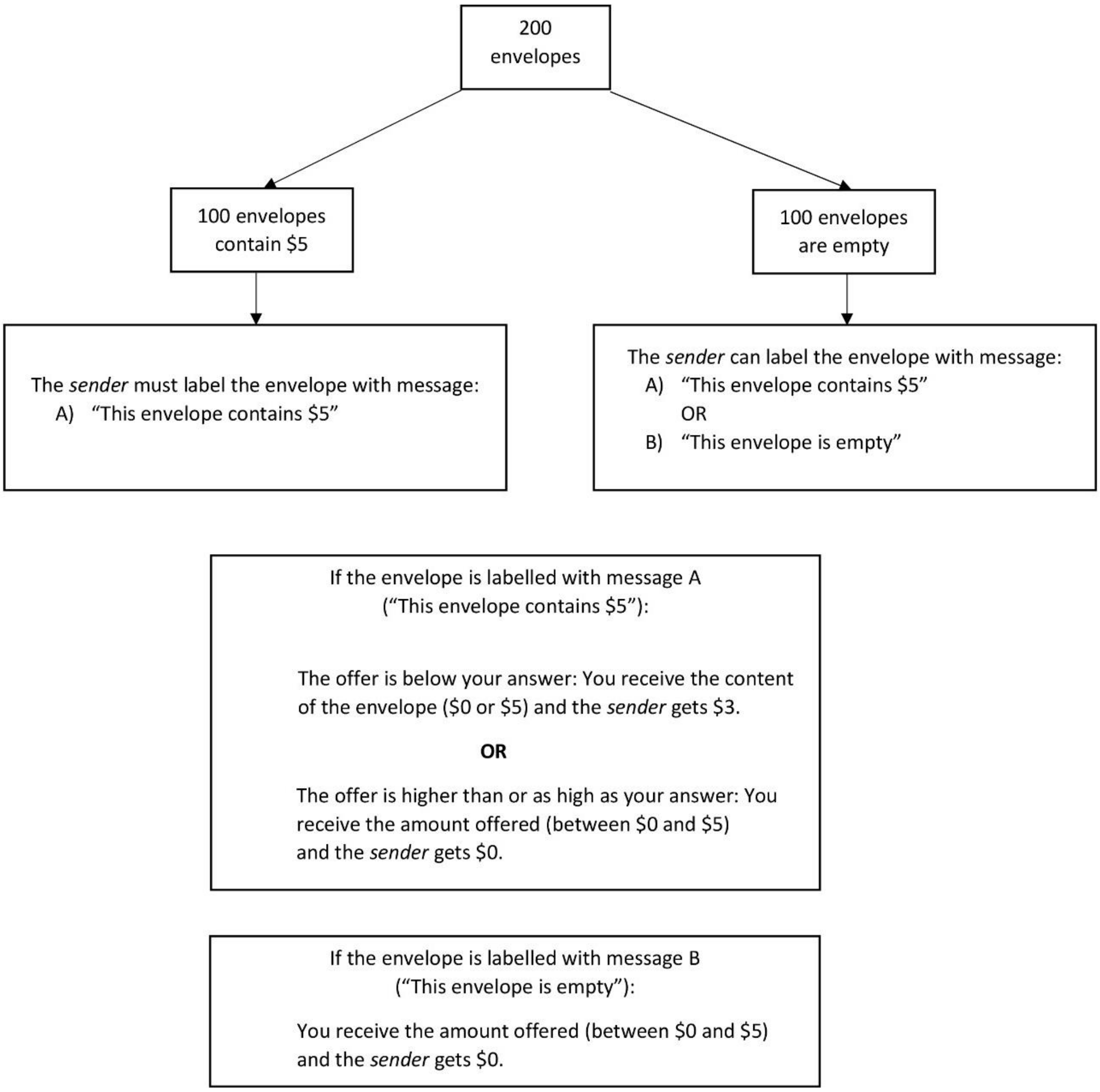
If the envelope is labelled with message B (“This envelope is empty”), you sell the envelope irrespective of the submitted *minimum selling price*.

- In this case you receive the offered amount, and the *sender* gets $0.

These instructions are provided to all participants of the experiment.

In a moment, we will ask you to submit a *minimum selling price*. The illustration on the next page summarizes the instructions above.

Task 1



The question you will be asked is:

If your envelope is labeled with message A ("This envelope contains $5"), for how much money would you be willing to sell it to us?

Comprehension Questions

Before we turn to Task 1, please answer the following comprehension questions. We will show you the correct solutions on the next page.

Question 1

How many of 200 envelopes contain $5?

Answer:

Question 2

How many of 200 envelopes are empty?

Answer:

Question 3

If the envelope is labeled with the message "This envelope contains $5", then it contains $5 <u>with certainty</u>.

This is:

Question 4

How much does the *sender* earn if you open the envelope?

Answer (in $):

Question 5

How much does the *sender* earn if you do not open the envelope?

Answer:

Question 6

If your envelope is labeled with the message "This envelope contains $5", then you will get the amount offered to you if the offer is <u>higher than or as high as</u> your minimum selling price.

This is:

Solutions

This page shows you the correct answers for the comprehension questions. Below, you can see whether your answer was correct and you can access additional information by clicking on „+ more info“.

Question 1

How many of the 200 envelopes contain $5? You answered: 100 envelopes
Your answer is correct

+ more info
„For this task we have prepared 200 envelopes in the following way: We put $5 in 100 envelopes. The remaining 100 envelopes remain empty, they thus contain $0."

Question 2

How many of the 200 envelopes are empty? You answered: 100 envelopes
Your answer is correct.

+ more info
„For this task we have prepared 200 envelopes in the following way: We put $5 in 100 envelopes. The remaining 100 envelopes remain empty, they thus contain $0.“

Question 3

If the envelope is labelled with message A ("This envelope contains $5"), then it contains $5 with certainty. You answered: *„False“*
The statement is false. Therefore, your answer is correct.

+ more info

"Each *sender* receives one randomly selected envelope and they are asked to label it with one of the two messages:

A. "This envelope contains $5"
B. "This envelope is empty"

The rules are the following: if the envelope contains $5, the *sender* can only choose message A. If, however, the envelope is empty, the *sender* can choose freely one of the two messages."

Question 4

How much does the *sender* earn if you open the envelope? You answered: $3
Your answer is correct

+ more info

"If the computer's offer is below the *minimum selling price*, you open the envelope. In this case you receive whatever amount the envelope contains, and the *sender* gets $3."

Question 5

How much does the *sender* earn if you do not open the envelope? You answered: $0
Your answer is correct

+ more info

"If the computer's offer is higher than or as high as the *minimum selling price*, you sell the unopened envelope. In this case you receive the offered amount, and the *sender* gets $0."

Question 6

If your envelope is labelled with message A ("This envelope contains $5"), then you get the amount offered to you if the offer is higher than or as high as your *minimum selling price*.
You answered: *„True“*
The statement is true. Therefore, your answer is correct.

+ more info

"If the envelope is labelled with message A ("This envelope contains $5"), you either open or sell the envelope:

- If the computer's offer is below the *minimum selling price*, you open the envelope. In this case you receive whatever amount the envelope contains, and the *sender* gets $3.
- If the computer's offer is higher than or as high as the *minimum selling price*, you sell the unopened envelope. In this case you receive the offered amount, and the *sender* gets $0."

Task 1

By clicking on the ‚+‘ symbols you can access the instructions again. Please take your time and consider your decision carefully.

+ Information on Task 1

+ Graphical Illustration

Additional information:

75% of all senders whose envelopes were empty, labeled them with the message A ("This envelope contains $5").

Below you can use a calculator that was designed specifically for this task. You can submit an estimate of the share of *senders* who labeled their empty envelopes with message A (“This envelope contains $5”). Based on your estimate, the calculator computes the probability that *your* envelope contains either $0 or $5. You can submit multiple estimates and use the calculator as often as you wish.

Input:

If % of those *senders* whose envelopes were empty, labeled them with the message A ("This envelope contains $5") and you receive an envelope that is labeled with message A, then:

- Your envelope contains $5 with a probability of %.
- Your envelope is empty with a probability of %.

Please answer the following question:

If your envelope is labeled with the message A ("This envelope contains $5"), for how much money would you be willing to sell it to us?

The amount can include up to two decimals for cents.

Your answer: $0.4

If we offer you between $0 and $0.39, you will open the envelope.

If we offer you $0.40 or more, you will sell the envelope.

Task 2

In this task you will receive a lottery ticket. The conditions of the lottery are explained in detail below. We will then ask you to name the lowest offer that you are willing to accept in exchange for the lottery ticket – the minimum selling price. The computer will then generate a random amount between $0 and $5 and offer it to you.
If the offer is below the minimum selling price, you will keep your ticket and participate in the lottery.
If the offer is higher than or as high as the minimum selling price, you will sell the lottery ticket in exchange for the offered amount.
However, with a probability of 12,5% you automatically sell the lottery ticket regardless of your answer.

The lottery in Task 2 can lead to earning for you and for one of the *senders* from Task 1. The conditions of the lottery are as follows:

Lottery

- With a probability of 57% you earn $5 and one *sender* earns $5. The *sender* in this case will be chosen randomly among all those whose envelope contained $5.

- With a probability of 42% you earn $0 and one *sender* earns $5. The *sender* in this case will be chosen randomly among all those whose envelope was empty and who chose to label it with the following message: "This envelope contains $5"

Please answer the following question:

For how much money would you be willing to sell us your lottery ticket?

The amount can include up to two decimals for cents.

Your answer: $

If we offer you between $0.00 and $, you will participate in the lottery.

If we offer you $ or more, you will sell the lottery ticket.

Task 3

In this task you will receive a lottery ticket. The conditions of the lottery are the same as in Task 2, except that now the *senders* do not have the chance to get a payment.
We will ask you to name the lowest offer that you are willing to accept in exchange for the lottery ticket – the minimum selling price. The computer will then generate a random amount between $0 and $5 and offer it to you.
If the offer is below the minimum selling price, you will keep your ticket and participate in the lottery.
If the offer is higher than or as high as the minimum selling price, you will sell the lottery ticket in exchange for the offered amount.

However, with a probability of 0,125% you automatically sell the lottery ticket regardless of your answer.

The conditions of the lottery are as follows:

Lottery

- With a probability of 57% you earn $5.

- With a probability of 42% you earn $0.

Please answer the following question:

For how much money would you be willing to sell us your lottery ticket?

The amount can include up to two decimals for cents.

Your answer: $

If we offer you between $0.00 and $ you will participate in the lottery.

If we offer you $ or more, you will sell the lottery ticket.

Task 4

This task consists of 2 parts. In the first part you must answer 3 questions and you earn $0.50 per correct answer.

Click "Next" to start with the first part of Task 4.

Task 4 - Part 1

Question 1:
A bat and a ball together cost $1.10. The bat is $1 more expensive than the ball. How much does the ball cost?

Answer (in cents):

Question 2:
5 machines take 5 minutes to produce 5 items. How long do 100 machines take to produce 100 items?

Answer (in minutes):

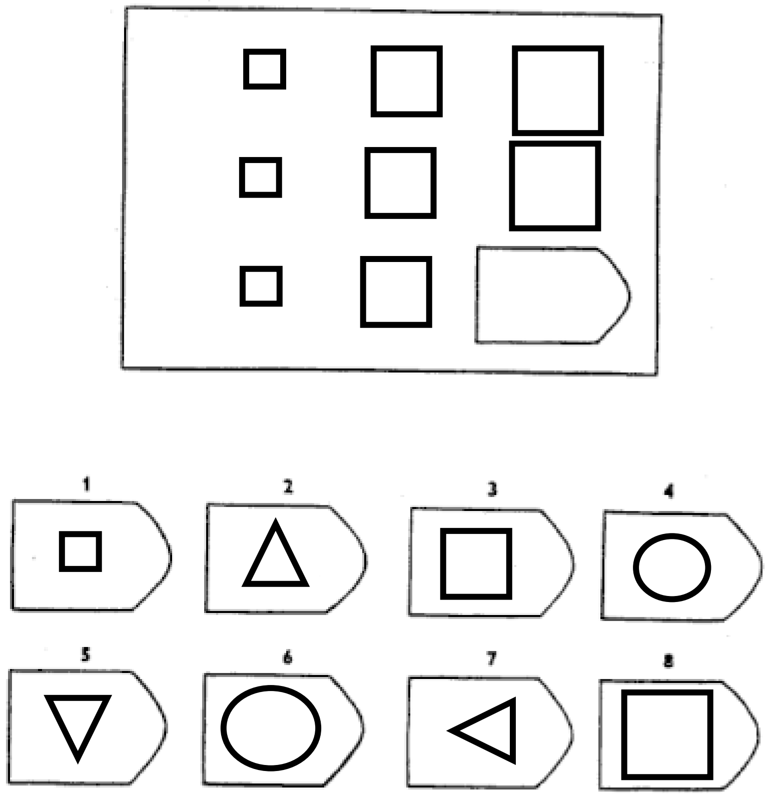
Question 3:
A lake is partially covered with lily pads. The area that is covered by lily pads doubles every day. If it takes 48 days for the lily pads to cover the whole lake, how long does it take until half of the lake is covered?

Answer (in days):

Task 4 - Part 2

In the second part of Task 4, we will show you a series of matrices containing patterns but with one missing piece. Your task is to pick the pattern that completes the matrix. For each matrix there is exactly one correct solution.

Here is an example:



Which one of these patterns completes the image?

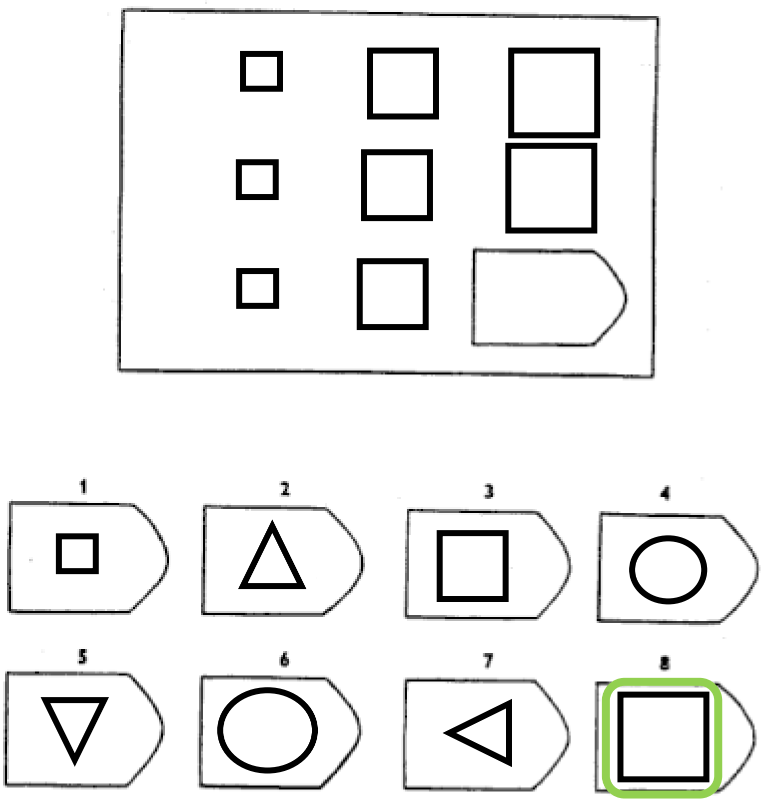
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

Please select one option and click on the "Next" button. You will see the correct solution on the next page.

Task 4 - Part 2

Your answer was correct.
Total number of correct answers: 1

The correct answer for the introductory example is pattern number 8.



Which one of these patterns completes the image?

- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

There are 34 matrices, and you have three minutes to solve as many as possible. You will earn $0.30 per correct answer. If Task 4 is chosen as the relevant one for your payoff, your earnings from part one and part two will be added and paid to you.

The problems are easy at the beginning, and gradually become more difficult. The timer will start when you click “Next”.

Additional Task A1

Please think back to Task 1, in which *senders* were asked to label an envelope with one of these two messages:

A. "This envelope contains $5"
B. "This envelope is empty"

100 senders had an envelope that was empty. How many of these 100 senders do you think chose message A?
Answer:

You will receive $1 if your answer deviates by less than three from the correct answer.

Additional Task A2

For each of the following rows, please choose either A or B. One of the rows will be randomly selected for payment at the end of the experiment.

Row		Option A	Option B	
1	○	100% chance of winning $0.05	75% chance of winning $1 25% chance of winning $0	○
2	○	100% chance of winning $0.10	75% chance of winning $1 25% chance of winning $0	○
3	○	100% chance of winning $0.15	75% chance of winning $1 25% chance of winning $0	○
4	○	100% chance of winning $0.20	75% chance of winning $1 25% chance of winning $0	○
5	○	100% chance of winning $0.25	75% chance of winning $1 25% chance of winning $0	○
6	○	100% chance of winning $0.30	75% chance of winning $1 25% chance of winning $0	○
7	○	100% chance of winning $0.35	75% chance of winning $1 25% chance of winning $0	○
8	○	100% chance of winning $0.40	75% chance of winning $1 25% chance of winning $0	○
9	○	100% chance of winning $0.45	75% chance of winning $1 25% chance of winning $0	○
10	○	100% chance of winning $0.50	75% chance of winning $1 25% chance of winning $0	○
11	○	100% chance of winning $0.55	75% chance of winning $1 25% chance of winning $0	○

12	○	100% chance of winning $0.60	75% chance of winning $1 25% chance of winning $0	○
13	○	100% chance of winning $0.65	75% chance of winning $1 25% chance of winning $0	○
14	○	100% chance of winning $0.70	75% chance of winning $1 25% chance of winning $0	○
15	○	100% chance of winning $0.75	75% chance of winning $1 25% chance of winning $0	○
16	○	100% chance of winning $0.80	75% chance of winning $1 25% chance of winning $0	○
17	○	100% chance of winning $0.85	75% chance of winning $1 25% chance of winning $0	○
18	○	100% chance of winning $0.90	75% chance of winning $1 25% chance of winning $0	○
19	○	100% chance of winning $0.95	75% chance of winning $1 25% chance of winning $0	○
20	○	100% chance of winning $1.00	75% chance of winning $1 25% chance of winning $0	○

Thank you for participating in today's experiment!

Before informing you about your earnings and the payment modalities we would like you to answer a few brief questions.

Final Questions

What is your gender?

- ○ Man
- ● Woman
- ○ Non-binary
- ○ Prefer not to say

What is your age?

What is your highest level of education?

Your Earnings

We will now break down your total earning:

1. For participating in the experiment, you earn $4.00.

2. Task 2 was randomly selected to be relevant. In Task 2 you stated that you were willing to sell your lottery ticket for $3.0 or more. The random device generated an offer of $0.95 for you.

   With a probability of 12.5% you <u>automatically</u> sell the lottery ticket. Random device 1 determined that this is not the case. The offer is therefore compared to your minimum selling price. Since it is lower than your minimum selling price, you play the lottery with a winning probability of 57%. Random device 2 determined that your lottery ticket did not win. You earn $0. A randomly chosen *sender* that labelled his or her empty envelope with message A (“This envelope contains $5”) earns $5.

3. In Additional Task A1 you answered that 58 senders of an empty envelope labelled it with message A (“This envelope contains $5”). The correct answer is 57. Your answer deviates by more than two from the correct answer, your earnings thus do not increase. / Your answer deviates by less than three from the correct answer, your earnings thus increase by $1.

4. In Additional Task A2, item 1 was chosen. You chose Option A. Your earnings increase by $0.05.

In today's experiment you earned a total of $6.05.

Chapter 3

Lax Rent Control - An Investigation of Rent Control in Vienna

3.1 Introduction

Rent regulations have been introduced in North America and Europe mainly during and after the First and Second World Wars, although some earlier examples can be found throughout history.[1] During wartime, relocation of labour creates pressure on local housing markets while in the aftermath the return of soldiers and a diminished housing stock can cause a disruptive rise in rents. Even though the appropriateness of wartime interventions seems to be undisputed, economists had reached a rare consensus by the end of the 20^{th} century about the negative effects of rent ceilings.[2] Nonetheless, rent regulations remain prevalent to this day.[3] Economists have since re-examined their opposition to such policies, as some of the potential negative effects have been theoretically and empirically disputed.[4] The evaluation of rent control policies is thus at least partly an empirical question.

[1]One of the earliest examples in modern history is the 1881 Land Act in Ireland. Historically, some forms of rent control can be found in the Papal States in the 16^{th} century.

[2]In a survey of economists' opinions, Alston et al. (1992) asked members of the American Economic Association whether they "generally agree", "agree with provisions", or "generally disagree" with 40 statements related to economic theory and policy. The greatest degree of consensus on any question - 93.5 percent — was agreement or qualified agreement with the statement:"A ceiling on rents reduces the quantity and quality of housing available."

[3]In a recent study, Kettunen and Ruonavaara (2021) find that 16 out of 33 analyzed European countries had some kind of rent control in place.

[4]Two the main concerns about rent control are reduced supply and maintenance. Sang Ho (1992) and Häckner and Nyberg (2000) show that the effect of rent control on housing supply is theoretically ambiguous, while Sims (2007) finds only a small effect empirically. Regarding maintenance, Olsen (1988) and Kutty (1996) offer theoretical results, while Moon and Stotsky (1993) and Arnott (1995) provide empirical evidence that the effect on maintenance is ambiguous.

From a political and welfare standpoint, the potentially negative indirect effects of rent control have to be put into relation with the direct effect of making housing more affordable for at least part of the population. The underlying premise, and one that has not been challenged in the previous literature, is that rent control does in fact lead to a reduction in rents in the regulated sector. In other words, studies of rent control policies typically assume that policies are complied with or enforced.

The aim of this paper is to analyse the effects of rent control policies under the aspect of imperfect enforcement. While previous studies have not explicitly taken into account issues of enforcement and compliance, this paper closes this gap in the literature. It does so by first formulating a model of rent control under imperfect enforcement, showing that such policies can have the unintended effect of increasing rents in the controlled sector. The reason for this is the way that a binding rent cap that is not complied with affects the demand and supply in the market. Demand increases since informed tenants, who can enforce the rent cap ex-post, demand housing according to the regulated rent rather than the posted rent. Supply decreases because landlords take into account the possibility of ex-post enforcement and supply housing according to the expected rent rather than the posted rent. These two effects lead to an increase in posted rents relative to the counterfactual where no rent control is in place. Furthermore, the lower the rent cap, the higher the posted rents. Strikingly, this increase is sufficiently large to offset the fact that the rent cap is enforced ex-post by informed tenants, leading to an increase in expected rents as well. Thus, the model predicts a *perverse* effect of lax rent control.

Motivated by these findings, I study the effects of rent control in Vienna, where compliance has been of concern in the political discourse because even though the Austrian rent control caps rents, it does not strictly enforce compliance.[5] Instead, landlords are initially free to set any rent which tenants can contest after having signed the rental contract. Importantly, having had the rent successfully contested, a landlord is not fined and does not have to bear any costs beyond reimbursing the tenant for the accrued overcharge. The result of these institutional details is that, given that the cap is binding and thus produces excess demand, it is a dominant strategy for landlords to charge rents above the legally allowed level. The institutional setup of the Austrian rent control closely aligns with the key features of the theoretical model, making it an ideal candidate for testing the model.

I start the empirical analysis by comparing posted rents to computed rent caps that are determined by the location and other features of the apartments. This comparison

[5] According to the expert opinion of professor Helmut Böhm on behalf of the Austrian Chamber of Labour the complexity and lack of transparency renders the regulation ineffective, see Rosifka and Postler (2010).

allows me to assess compliance with the Austrian rent control regulation. Next, I use two institutional details of the Austrian rent control (*Mietrechtsgesetz*, MRG) to estimate its effects on rents. The first detail involves the applicability of the rent cap, which only applies to apartments between 30 and 130 m^2. This enables me to employ a regression discontinuity design to estimate the effect of the regulation on posted rents. The second detail involves the level of the rent cap, which varies across the city due to location premiums defined at the level of census districts. By leveraging the granularity of these districts and the discrete jumps in location premiums, I use a spatial discontinuity design to estimate the effect of a change in the rent cap on posted rents.

My findings lend partial support to the hypotheses generated by the model. I provide evidence that compliance rates are indeed low across the entire city with about 86% of the posted rents violating the legal rent caps. The estimated local average treatment effect (LATE) around the threshold of $130m^2$ is insignificant, implying that unregulated apartments just above the threshold are not more expensive on average than their regulated counterparts just below the threshold. The estimated average treatment effect of an increase in the rent cap is positive but not highly significant, implying that a 1€ higher rent cap increases posted rents by about 0.25€ .

Overall, the findings highlight the importance of proper enforcement of rent control policies even though they are ambiguous with regard to the predictions of the theoretical model. Further research is necessary to provide a more conclusive understanding of imperfectly enforced rent control policies.

Although the MRG has specific features that can lead to overcharging, non-compliance is a common concern with rent controls and price regulation in general. For example, in 2016, the German newspaper *Zeit* reported that the German Tenants' Association found that between 67% and 95% of rental offers in Berlin, Hamburg, Munich, and Frankfurt am Main exceeded the permitted upper limits.[6] Similarly, Lemos (2009) documents non-compliance with minimum wage laws in Brazil. Therefore, the findings presented in this paper are likely to have relevance for other price-controlled markets as well.

The remainder of the paper is structured as follows: Section 3.2 describes the paper's contribution to the related literature. Section 3.3 provides details about the institutional background of the MRG and stylised facts about the Viennese rental market. A theoretical model of imperfectly enforced rent control is provided in Section 3.4. Section 3.5 describes the data. Section 3.6 lays out the empirical strategy to test the hypotheses generated by the model and presents the findings. Further steps are outlined in Section 3.7 and Section 3.8 concludes.

[6] https://www.zeit.de/wirtschaft/2016-09/mieterbund-mietpreisbremse-vermieter-ignorieren-reform?utm_referrer=https%3A%2F%2Fpdf.sciencedirectassets.com%2F

3.2 Related Lierature

This paper is related to the rich literature on rent control of which I do not attempt to provide an overview. I refer to Jenkins (2009) for a survey of the literature as well to Kholodilin (2022) for a more recent survey of the empirical literature on rent control and instead focus on the point of my departure. The literature on rent control typically assumes that rent control, whichever specific form it takes, is complied with. Contrary to that, this paper studies the phenomenon of non-compliance. It, therefore, relates to the literature on weakly enforced price controls, which has so far been studied mainly in the labour literature. Ashenfelter and Smith (1979) and Lemos (2009) provide empirical evidence for non-compliance with minimum wage laws while Basu, Chau and Kanbur (2010) and Danziger (2010) examine the issue from a theoretical standpoint. Basu et al. (2010) study a monopsonistic employer and give conditions under which he will comply with a wage regulation even if he is only checked with some probability while Danziger (2010) shows that in a competitive market, small firms react to such weak regulations by employing fewer workers and paying them less. I adapt the model in Basu et al. (2010) to show that in a competitive rental market, an imperfectly enforced rent control can have similar perverse effects. The empirical part of the paper provides empirical evidence for non-compliance with the MRG and estimates the effect on rents of this imperfectly enforced rent control policy.

I further contribute to the literature by providing the first causal analysis of the effect of rent control in the context of the Austrian MRG. The best evidence so far comes from analyses of German rent control policies. In 2015 Germany's Federal Government introduced a rent control (*Mietpreisbremse*) similar to the MRG. Like the Austrian regulation, it controlled both initial rent levels of new contracts and increases over time within a tenancy.[7] Likewise, due to the complexity of the regulation and lack of enforcement, concerns about its effectiveness arose. Empirical studies of German rent control thus provide a useful benchmark for the Austrian case, not least because the German and Austrian rental markets have a similar tenancy structure. Using a difference-in-difference approach, Mense, Michelsen and Kholodilin (2018) estimate a significant decrease in rent levels for *de facto* regulated dwellings following the introduction of the rent regulation.[8] In a more recent study, Breidenbach et al. (2022) apply a triple difference event study approach to analyse temporal dynamics of the rent control to find that

[7]The institutional details differ. The German regulation caps new rental contracts at 10% above the comparative rent index. Increases are allowed to up to 20% over three years as long as the resulting rent does not exceed the comparative rent index (*Mietspiegel*).

[8]They split the treated sample into a *de facto* treated group that experienced high rent growth prior to the regulation and a *de facto* untreated group that did not.

while the initial rent dampening effect is significant, it vanishes about one year after the implementation.[9] Most recently, Hahn, Kholodilin, Waltl and Fongoni (2023) study the short-run effects of a rent freeze in Berlin that was introduced in February 2020 by the federal state of Berlin and abolished 13 months later by the German constitutional court. Their analysis relies on the fact that the regulation applied only within the city border while Berlin's urban region reaches into the surrounding federal state of Brandenburg. The resulting discontinuity at the city border is used to identify the causal effect of the rent freeze.

Hahn et al. (2023) is therefore methodologically most closely related to the present study. The main difference is that instead of a discontinuity along a long, circular border I exploit more granular variation within the city of Vienna. Moreover, while Hahn et al. (2023) study the short-term effects of a policy that has been in place only briefly, my analysis investigates the long-term effects of a policy that has been in place for several decades. The granular approach in my paper is similar to Giacomelli and Menon (2017), who exploit spatial discontinuities in court jurisdictions in Italy to study how contract enforcement affects firm growth.

3.3 Institutional Background and Stylised Facts

The various forms that rent control can take have been broadly categorized into *first*, *second*, and *third-generation* rent control in the literature.[10] While first-generation rent control imposes an absolute cap on rents, controls of the second generation limit the increases of rents over time. Typically, only the existing housing stock is regulated while new constructions remain uncontrolled. Rent regulation policies that control rents within a tenancy but not between tenancies are called third-generation rent control.

These categories are of course very crude and rent regulations have many more aspects according to which they can differ from one another. Nevertheless, they offer a useful framework for comparing rent control policies.

The Austrian rent regulation (*Mietrechtsgesetz*, MRG) has elements of both first and second-generation policies. On the one hand, it imposes a hard rent cap on apartments in historical buildings which, however, adjusts to the specifics of an apartment and its location. This aspect is reminiscent of first-generation rent controls. On the other hand, these rent caps are regularly adjusted according to the consumer price index (CPI),

[9]The regulation's staggered implementation allows Breidenbach et al. (2022) to define two control groups, one defined by dwelling characteristics (old vs. new) and one defined via location (implemented vs. not implemented).

[10]See, for example, Kettunen and Ruonavaara (2021).

allowing for increases over time while regulating the extent of it. Thus, the MRG also contains aspects of second-generation rent controls.

3.3.1 Institutional Background of the MRG

Rent regulation in Austria goes back to the early 20^{th} century. Necessitated by urban population growth due to industrialisation, the first comprehensive set of housing laws was passed after WW1 and included things like protection against dismissal, restrictions on contract terms, rent caps, and rights for claiming past overcharges. This first-generation rent control quickly became unfeasible when rapid inflation let the absolute rent cap go to almost zero in real terms. Several reforms followed of which the most significant was passed in 1929 which rigorously defined scopes of application for the first time. The first complete reformulation of the housing law took place in 1981 when the "Mietrechtsgesetz" (MRG) was passed. It uncluttered the historically grown landscape of regulations and introduced so-called category rents ("Kategoriemietzins") which defined rent caps based on explicitly defined apartment characteristics for all of Austria. While clear and transparent, these (regionally) uniform rent caps failed to meet the realities of the housing market, where location and amenities crucially affect prices. After a number of subsequent reforms Austria's housing law is once again a rather convoluted set of rules and regulations.

The current regulation is quite complex because rather than being a simple uniform price cap it depends on several hard and soft characteristics of the apartment like size, features and facilities, location, and contract length. On top of that, there are numerous exceptions where the regulation doesn't apply and the rent can be set freely. Generally, the price cap applies to rental items in houses built before 1945. However, exceptions include:

- houses that have been rebuilt after WW2 but where the government loans used for rebuilding have been repaid prematurely
- rental objects resulting from loft conversions if the corresponding building permit was issued after 2001
- rental objects in houses that are subject to protection of historic monuments
- rental objects in houses with only 2 or fewer rental objects
- rental objects with a size below $30\,m^2$ or above $130\,m^2$

At the same time, rental items are sometimes subject to the cap even if they are not built before WW2, namely if:

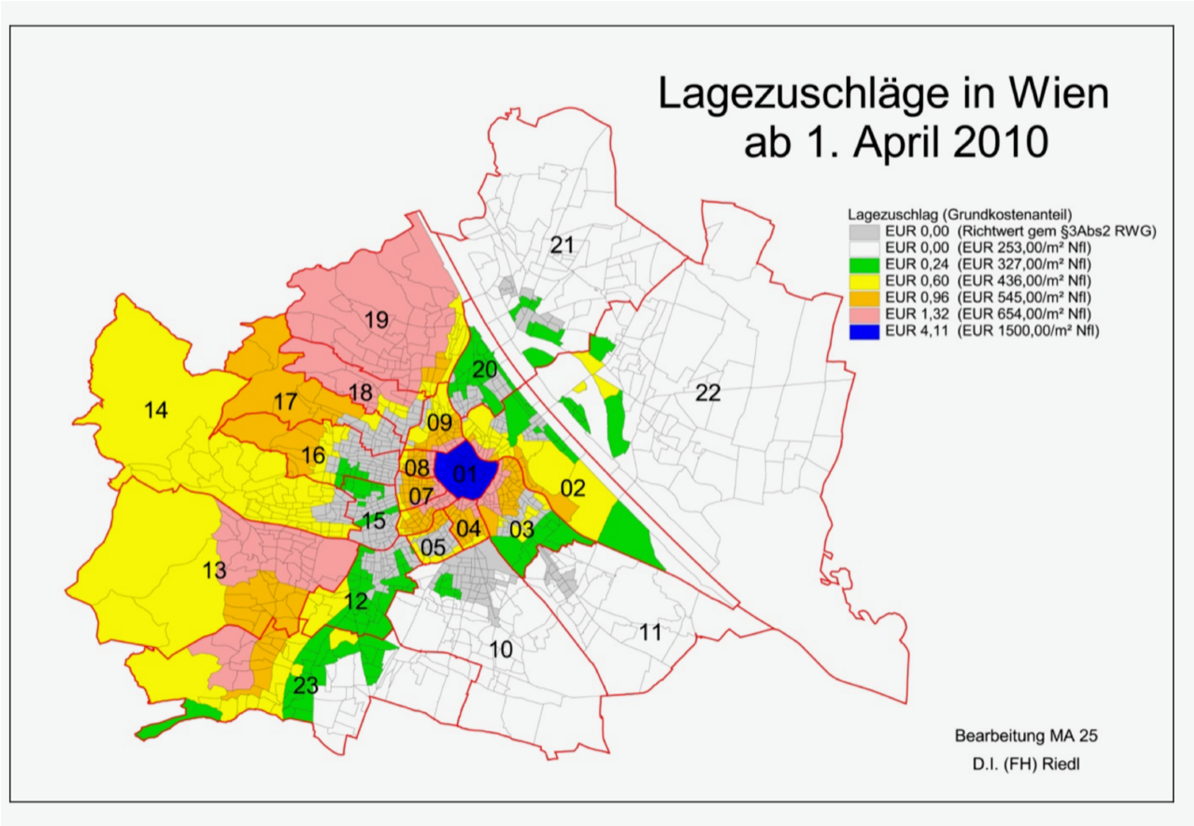


Figure 3.1: The maps show the distribution of location premiums across the city of Vienna. The height of the different premiums and the census districts to which they apply are determined my the municipal government.

- the house has been built before 1953 using a government loan (but only if that loan has not been repaid prematurely)
- the rental object was built after 1945 under the subsidised housing program

If the rental contract is limited - regardless of the length of the contract term - the price cap reduces by 25%. The way location factors in is similarly complex. Different surcharges are justified depending on both the location of the house within the city and the apartment's location within the house. Location premiums are defined as discrete surcharges that apply to regulated rental objects within a census district (Zählsprengel, ZSP). Generally, several ZSPs are collected to define a location premium zone, although isolated ZSP can be found as well. These premiums are set by the local authority and published in the form of maps like the one shown in Figure 3.1. Roughly every two years an updated map with inflation-adjusted premiums is published. Less frequently, location premium zones shift, resulting in a variation in location premiums in space and over time.

3.3.2 Stylized Facts about the Viennese Rental Market

Renting makes up a large part of the Austrian housing market at around 40% and is particularly important in Vienna, where 75% of households live in rental apartments (Baumgartner, 2013). Moreover, the MRG is most important in Vienna, where around 500.000 apartments are subject to the rent cap (including ca. 200.000 council apartments which the city owns and manages, see Pichlmair, 2013) Hence, even though the MRG is a federal law it mainly affects the Viennese rental market.[11]

Among comparable cities worldwide, Vienna is considered to have relatively low costs of living. This is largely due to the public authorities' continued efforts to keep rents at a moderate level.[12] Nevertheless, housing costs in Austria have increased substantially in recent years. According to a report by the Austrian Institute of Economic Research, rents have increased by 23.9% between 2005 and 2012 (Baumgartner, 2013). Interestingly, rents in historical buildings have experienced a larger increase than those in non-historical buildings (ca. 30% vs. ca. 20%). This is despite the fact that rent caps have increased by only 12% during that period, showing the disconnect between the rent regulation and the very rents it is supposed to regulate.[13]

Following this sharp and mostly illegal increase in rents in historical buildings, several legal financing companies entered the market beginning in 2013.[14] The logical consequence of their market entry and marketing efforts is that more people are aware of rent overcharging. Arguably, over time this share also naturally increases as people learn about it by word of mouth. Even more importantly, third-party legal financing has made it easier for people to contest overpriced rents *conditional on being aware*. This is because disputes often (but not always) have to be settled in court and even though an overwhelming majority of such cases are judged in favour of the tenant, the risk associated with litigation is likely to be a deterrence.[15] Third-party legal funding eliminates this risk and is thus likely to increase the probability that a tenant challenges his or her rent if it exceeds the legal rent cap. A 2009 study by the Institute of Empirical Social Research (*Institut für empirische Sozialforschung*, IFES) on behalf of the Austrian Chamber of Labour finds that roughly half of the interviewed people did not know about rent regulations. The

[11]Another reason is that the rent caps vary by region and only in Vienna are they significantly below market rent and thus constitute a severe market intervention.

[12]In addition to the rent regulation, Vienna's city government invests heavily in subsidised housing with yearly spending averaging over 500 million €.

[13]See https://www.mieterschutzwien.at/65/richtwerte for the evolution of rent caps.

[14]Legal financing is the service of providing third-party funding for legal disputes. Typically, the business model is based on commissions in form of percentage shares of the judgment or settlement. mieterunter.at, mietheld.at, mietfuchs.at and mietenchecker.at are the four of the main companies offering such services for housing disputes in Vienna.

[15]According to Rosifka and Postler (2010) more than 98% of cases are settled in favour of the tenant.

same study found that only 6% had their contracts checked for overcharge while another 21% were considering such a step (Feistritzer, 2010). This is particularly interesting given the fact that by and large, landlords do not comply with the rent caps, as my empirical analysis in Section 7 will reveal.

3.4 A Model of Imperfectly Enforced Rent Control

The starting point for the analysis is a simple theoretical model that does not attempt at describing the residential rental market in a complete and realistic way. Rather, it offers a simple framework that allows us to analyse imperfectly enforced rent control in a tractable fashion. There is a rental market with two sectors: a regulated sector that is potentially subject to rent control, and an unregulated sector. I will first derive the equilibrium in the benchmark case *without any* rent control and look at the effects of rent control in the next section.

There is a mass of tenants and they differ in their preference for renting an apartment in a historical building.[16] Tenant i assigns valuation t_i to living in a historical building (instead of in a non-historical one). Denote by the rent by r and let $u_T(r,t) = r - t$ represent a tenant's *disutility* of renting a historical apartment. It is given by the rent level and reduced by the extent of her preference for historical buildings. Compared to this, the disutility of living in a non-historical building is given by a, the rent level in the non-historical segment.[17] Assuming that preferences for historical buildings are uniformly distributed on $[0, T]$, we can interpret the parameter T as the relative attractiveness of historical buildings relative to non-historical ones.[18]

To derive demand in the historical segment, note that a tenant will rent an apartment in the historical segment if his disutility of renting it, u_t, is smaller than his disutility of renting in the non-historical sector, a. That is, the probability that tenant i will rent an apartment in the historical segment is given by $Pr(r - t_i < a)$ which yields the following demand:

$$h^d(r) = 1 - \frac{r-a}{T} \tag{3.1}$$

[16]In the private sector of the rental market, most of the historical buildings are subject to rent control and most of the controlled apartments are in historical buildings. For simplicity, I treat the historical sector as the regulated one.

[17]I take the rent level in the non-historical (non-regulated) segment as given and think about it simply as a given outside option. Arguably, the rent levels in different segments depend positively on each other. Considering this dependence only strengthens our results.

[18]One way to think about the support $[0, T]$ is that every tenant likes historical buildings at least as much as non-historical buildings. Alternatively, one can think about it as a normalization of a more general interval.

As expected, it decreases with the rent level in the (potentially) regulated sector, r, and increases with the rent level in the unregulated sector, a. Moreover, demand goes up as historical housing becomes relatively more attractive, i.e. as T increases.

On the other side of the market, there is a mass of landlords, each possessing a historical apartment. They can either rent it out at rent level r or sell it on the real estate market for a price of b.[19] Heterogeneity among the landlords is captured by s_j. One can think of s_j as capturing how skillful landlord j is at all the things that come with renting out an apartment or capturing that landlord's preferences for owning a historical apartment. Both of these aspects affect a landlord's utility of renting out his apartment instead of selling it. Adopting the latter interpretation, I let $u_L(r,s) = r + s$ represent a landlord's utility from renting out a historical apartment. Since his outside option is selling it at price b, he will rent whenever $r + s > b$. Assuming that landlords' preferences are uniformly distributed over $[0, S]$, we can derive rental housing supply in the historical segment.

$$h^s(r) = 1 + \frac{r-b}{S} \tag{3.2}$$

It depends positively on the rent level, and negatively on the price of housing in the real estate market. An increase in S, which can be interpreted as a stronger preference of landlords to keep ownership of historical apartments by renting them instead of selling, increases supply.

From equations (3.1) and (3.2), we can derive the equilibrium in a market without rent control (solid lines in Fig. 3.2). The equilibrium levels of rent and housing in this market are given by

$$r^* = \frac{aS + bT}{S+T} \qquad\qquad h^* = 1 + \frac{a-b}{S+T} \tag{3.3}$$

Recall that a and b are the rent level in the unregulated sector and the apartment price in the real estate market, respectively, and let us examine their relationship with r^* and h^*. First, consider an increase in the rent level in the unregulated sector. When a goes up, tenants' outside option becomes less attractive so they demand more housing in the regulated sector. This drives up the rent in that sector so more landlords are willing to supply housing. Consequently, equilibrium levels of rent and housing both increase. This intuition is in line with Equations (3.3), where r^* and h^* depend positively on a. Next, consider an increase in the apartment price in the real estate market. When b goes up, landlords' outside option becomes more attractive so they supply less housing in the regulated sector. This, in turn, increases the rent in that sector. Thus, the equilibrium quantity of housing decreases, and the equilibrium rent level increases. This intuition is

[19]To meaningfully compare r and b, we should think of r as the *net present value of renting at rent r.*

also in line with Equations (3.3), where r^* depends positively and h^* depends negatively on b.

3.4.1 Rent Control

Now let us introduce a rent regulation policy that regulates rents in the historical segment. This policy introduces three additional parameters into our model: $\bar{r}, \lambda,$ and σ.

The rent cap is denoted by $\bar{r}$ and I assume that it is binding, i.e., that it is lower than the rent level in the benchmark case without regulation, $\bar{r} < r^*$.

I further assume that only a fraction λ of tenants are aware of the rent control. The remaining fraction of tenants believes that the market is unregulated and that they will have to pay the posted rent.

The third parameter, σ, captures frictions in the enforcement of rent control. In the simplest case, one might think of σ as a settlement cost, such that when the rent cap is enforced only a fraction $(1-\sigma)$ of the settlement is reimbursed. Such settlement costs arise in the presence of legal fees. Another way to think about σ is that it captures the uncertainty that the convoluted regulation creates. Because the requirements for when the cap applies are complex, a tenant is not certain that a claim will be successful. Following this interpretation, σ represents the probability that a claim will be successful. Below, I will use the first interpretation.

The government does not enforce the regulation actively, nor does it sanction it. The only way to have it enforced is for a tenant to actively submit a claim. If it goes through the tenant is reimbursed for the overcharge, i.e. she receives a transfer $r - \bar{r}$ from her landlord but she bears the settlement cost $\sigma(r - \bar{r})$. Other than having to reimburse the tenants, there is no fine for violating the rent cap. Therefore, posting a rent above $\bar{r}$ is a weakly dominant strategy for landlords.

Such a rent policy affects the market in the following way. Uninformed tenants demand housing according to posted rents, not knowing that a lower rent can be enforced. Their demand function is therefore not affected by the policy. Informed tenants also observe the posted rent but anticipate receiving a reimbursement of (1-σ)$(r - \bar{r})$ after submitting a claim. They thus demand housing according to the *effective rent* $r - (1-\sigma)(r - \bar{r})$. Denoting by $\bar{h}_u^d(r, \bar{r})$ the housing demand of uninformed tenants in a market with rent cap $\bar{r}$ and by $\bar{h}_u^d(r, \bar{r})$ that of informed tenants, we have:

$$\bar{h}_u^d(r, \bar{r}) = 1 - \frac{r-a}{T} = h^d(r) \tag{3.4}$$

$$\bar{h}_i^d(r, \bar{r}) = 1 - \frac{r - (1-\sigma)(r-\bar{r}) - a}{T} = h^d(r) + (1-\sigma)(h^d(\bar{r}) - h^d(r)) \tag{3.5}$$

Given a posted rent r, a rent cap $\bar{r}$, settlement probability $1-\sigma$, and a share of

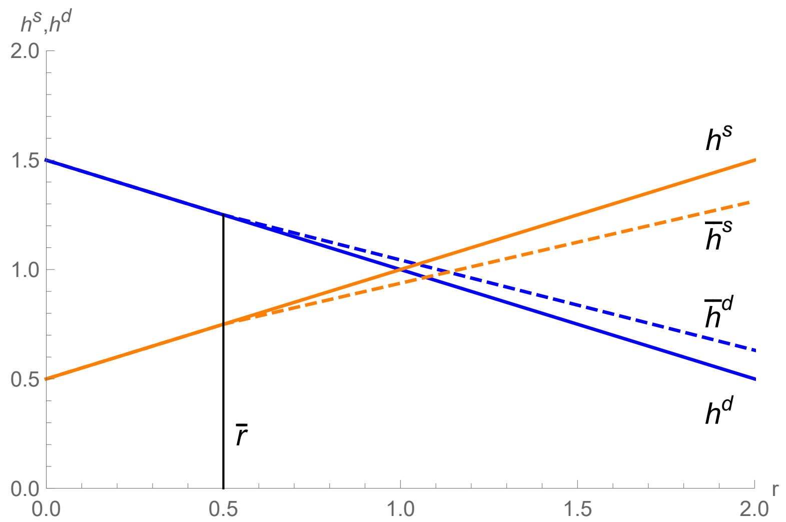


Figure 3.2: Equilibrium in unregulated (solid) and regulated (dashed) case. Housing amount decreases and rent level increases due to regulation.

informed tenants λ, total housing demand in the regulated market is given by

$$\bar{h}^d(r,\bar{r}) = (1-\lambda)\bar{h}^d_u(r,\bar{r}) + \lambda\bar{h}^d_i(r,\bar{r}) = h^d(r) + \underbrace{\lambda(1-\sigma)(h^d(\bar{r}) - h^d(r))}_{\text{ficticious demand}} \tag{3.6}$$

where the fictitious demand stems from the circumstance that some tenants expect to pay less than the market rent. Thus, from a landlord's perspective, it is fictitious in the sense that it doesn't reflect a response to the posted rent. Graphically, lax rent control rotates the housing demand counterclockwise around $(\bar{r}, h^d(\bar{r}))$.

The housing supply is affected because landlords rationally expect to earn the posted rent only with a probability of λ. Denoting by $\bar{h}^s(r,\bar{r})$ the supply of housing in a regulated market with an imperfectly enforced rent cap $\bar{r}$, we have:

$$\bar{h}^s(r,\bar{r}) = 1 + \frac{(1-\lambda)r + \lambda\bar{r} - b}{S} = 1 + \frac{r-b}{S} - \lambda\frac{r-\bar{r}}{S} = h^s(r) - \underbrace{\lambda(h^s(r) - h^s(\bar{r}))}_{\text{decrease in supply}} \tag{3.7}$$

For every $r > \bar{r}$, the supply in the regulated market is below the benchmark without rent control. Graphically, the supply is rotated clockwise around $(\bar{r}, h^s(\bar{r}))$. Put together, Equations (3.6) and (3.7) form an equilibrium in the regulated market, which yields the *market rent in the regulated market*, $\bar{r}^*$, and the corresponding level of housing, $\bar{h}^*$:

$$\bar{r}^* = \frac{aS + bT - \lambda[T + (1-\sigma)S]\bar{r}}{(1-\lambda)(S+T) + \lambda\sigma S} \qquad \bar{h}^* = 1 + \frac{(1-\lambda)(a-b)}{(1-\lambda)(S+T) + \lambda\sigma S} \tag{3.8}$$

Figure 3.2 depicts the market equilibria in the unregulated and regulated cases, respectively. As can be seen, under lax rent control, the market rent goes up and the amount of housing goes down compared to the unregulated benchmark case.

Despite the market rent going up, the rent regulation could still be effective if rents went down *in expectation*. If enough tenants paid only the regulated rent, then the expected rent would go down despite the increase in posted rents. This, however, turns out not to be the case. Some algebra yields that

$$\mathbb{E}[r] = \lambda(1-\sigma)\bar{r} + [1-\lambda(1-\sigma)]\bar{r}^* > r^* \tag{3.9}$$

Thus, not only do tenants get driven out of the market (housing amount decreases) but those who remain in the market end up paying more in expectation. Plugging in $\bar{r}^* = \frac{aS+bT-\lambda[T+(1-\sigma)S]\bar{r}}{(1-\lambda)(S+T)+\lambda\sigma S}$ into (3.9) yields

$$\mathbb{E}[r] = \frac{[1-\lambda(1-\sigma)](aS+bT) - \lambda\sigma T\bar{r}}{[1-\lambda(1-\sigma)](S+T) - \lambda\sigma T} \tag{3.10}$$

Examining Equation (3.10) reveals that the expected rent increases as the rent cap, $\bar{r}$, decreases. It also increases as the share of informed tenants, λ, or the settlement cost, σ, increases. This allows us to formulate the following hypotheses:

H1: Under lax rent control, landlords set rents above the rent cap.

H2: Under lax rent control, landlords set higher rents than without rent control.

H3: Under lax rent control, as the rent cap increases the rents set by landlords decrease.

3.5 Data

For this project, I draw from three different sources of data. First, data on rental listings are generously provided by the Research Unit Urban and Regional Research at the TU Wien. The data stem from several different real estate platforms and were compiled by Professor Wolfgang Feilmayr. Second, data on apartment-specific surcharges and deductions were obtained from the official website of the Viennese government.[20] Lastly, data on location premiums were extracted from corresponding maps using a geographic information system (GIS) software. Below, I provide details about each individual part of the data, their sources, and the manipulations done on my side.

3.5.1 Rent Data

The biggest part of the data for this project consists of listings of rental objects, which Professor Wolfgang Feilmayr compiled from several different real estate platforms. It contains a large set of apartment-specific attributes including variables like `rent`, `size`, `floor` (level), `no. of bathrooms`, etc., allowing us to control for most factors that are commonly understood to impact housing prices and rents. The raw dataset initially

[20] https://mein.wien.gv.at/Meine-Amtswege/richtwert?subpage=/lagezuschlag/

consisted of 149,593 listings from the period 2005 to 2017 a substantial part of which were duplicate observations. Duplicates occur either because a listing is published on multiple platforms, as is often the case, or because a listing gets re-published over time. To eliminate such duplicates, I drop all observations that differ at most in their time of publishing. This eliminates approximately half of the observations. Next, since I have to match every apartment with the corresponding location premium, for which I only have data from 2010 onward, I restrict the sample to the period after and including 2010. Finally, in order to restrict our attention to the segment of the market that is subject to the MRG, I keep only apartments in buildings that were constructed before and including 1945. While this is rather restrictive, I want to make sure to have as few "false positives" (apartments that are not subject to the MRG but I keep nevertheless) as possible. This comes at the cost of potentially many "false negatives", as buildings from the period after 1945, and in particular between 1945 and 1954, are often subject to the MRG as well.

Missing values are dealt with in a few different ways. The four crucial variables are `size`, `rent`, `location`, and `time of listing`. The first two are used to compute the per m^2 rent while the latter two are needed for merging the rental listings with the location premium data. Since I cannot make meaningful inferences or educated guesses, I drop observations with missing values for any of these variables from the dataset. I also drop observations that do not have an unambiguous value for `floor` or that are on the attic level. The reason is that apartments on the attic level are exempt from rent caps in many cases.[21]

For the rest of the variables, I make assumptions about missing values. For `garage`, `central heating`, `basement`, and `lift`, I assume that apartments have those features only if clearly indicated. Missing values are interpreted as 0, i.e. as not having the corresponding feature. For `# of balconies`, `# of loggias`, `# of terraces`, Iassume 0 unless indicated otherwise. I further assume that `# of bathrooms`, and `# of toilets` equals 1 if not indicated otherwise. Note that these assumptions are conservative with respect to our estimate of the share of non-compliant listings since any of these features increases the apartment-specific rent cap (see the Section 3.5.2).

I eliminate apartments that are smaller than $30\,m^2$ and larger than $300\,m^2$. Apartments below $30\,m^2$ are exempt from the rent control but because only a few apartments are this small, I cannot exploit the discontinuity for the RD design. I do take advantage of the upper threshold - apartments larger than $130\,m^2$ are also exempt - for the RD analysis.

[21]In particular, they are exempt if the conversion into an apartment took place after 2001 even if the building itself is older. Since I only observe the construction year of the building and not that of individual apartments, I cannot rule out such cases individually. Moreover, attic conversions typically result in luxurious apartments which is why they pose a severe threat of bias.

However, I eliminate rental objects with a size of more than $300\,m^2$ since they are very likely to be unusually luxurious apartments or intended for commercial purposes. The increased variance of the per m^2 rent for these large apartments supports this concern.

Lastly, I drop listings in the first district for the purpose of my empirical analysis. Being the historical city center, the first district has a high share of buildings that are protected as historic monuments and therefore exempt from rent control. Furthermore, rent and real estate prices in this district are much higher than in the rest of the city, which is both due to supply side (apartments are on average larger and more luxurious) and demand side (the historical city center is prestigious and in high demand also from foreign investors) differences. The raw data together with the Python code for the described cleaning process is available upon request and produces a final dataset of almost $6,000$ observations.

3.5.2 Apartment-specific Surcharges and Deductions

As described in section 3.3.1, the rent cap that applies to a particular apartment depends, among other things, on the features and characteristics of that apartment. For Vienna, the municipal authority defines these surcharges and deductions.[22] Table 3.1 shows an exhaustive list of 21 features with their associated surcharges and deductions as obtained from the official website of the municipal governement.[23] The seven features in the first column (`in teletype`) we observe in our data. The first three features in the second column (*in italic*) are standard features which I assume most apartments have. The remaining features are unobserved but vary between apartments. For these, we will consider three scenarios, an optimistic one where we give every apartment 'the benefit of the doubt' and assume it has all features which allow for additional premiums and none of the ones that lead to deductions, a pessimistic one where we assume the opposite and a realistic one where we assume it has 'reasonably good' features. More precisely, we assume for the realistic scenario that apartments have features that are marked with an asterisk in Table 3.1. The optimistic scenario allows me to show that rent cap violations cannot be explained by unobserved apartment features.

Another major deduction depends on a characteristic of the rental contract which I do not observe. Fixed-term contracts, irrespective of the term length, are subject to a 25% deduction. Because rental contracts are frequently fixed term the rent cap that I compute is lenient and our estimate of the compliance rate is going to be biased towards 100%.[24]

[22]The responsible authority in Vienna is the *Magistratsabteilung 25*

[23]https://mein.wien.gv.at/Meine-Amtswege/Richtwert/

[24]According to the Austrian microcensus data, 70% of rental contracts in Austria in 2017 had a fixed contract term. This is an increase from 58% in 2008.

Table 3.1: additional surcharges and deductions (% of base rent cap)

`elevator` (+9.09%)	*communal TV antenna (+0.73%)*	quiet location* (+10%)
`garden` (+10%)	*telephone connection (+2%)*	[very] noisy location (−[15]7%)
`balcony` (+2.5%)	*interphone* (+1%)	sloping ceilings (−5%)
`terrace/loggia`(+5%)	central heating (+3.55%)	extraordinary view (+5%)
2^{nd} `bath` (+5%)	bicycle/stroller room* (+1.14%)	no separate toilet (−2.5%)
2^{nd} `toilet` (+5%)	communal laundry room* (+1.82)	kitchen in hallway (−5%)
`floor` (+1.5% per floor)	communal room (+1.32%)	freshly renovated* (+5%)

3.5.3 Location Premiums

As described in section 3.3.1, the rent cap that applies to an apartment depends on its geographic location within the city via the location premiums defined by the municipal authorities. These premiums come in discrete categories and are adjusted according to a CPI in irregular intervals of 1 to 4 years. Besides adjustments to the rate of inflation, sometimes regions 'switch categories', such that the premium decreases or increases beyond the common trend. While I observe category switching in the data, it is not very common.

In principle, these data are accessible via the official website of the municipal government. However, as confirmed by the responsible authority, the location premiums before 2017 do not correspond to those that actually applied at the time. This is because in 2017 a new formula was introduced to compute location premiums and this formula was applied retroactively to previous periods. Therefore, I instead rely on graphical representations that are typically published whenever adjustments take place. Figure 3.1 shows such a map for the year 2010. I was able to obtain such maps for all adjustments in the period between 2010 and 2017 and used the *QGIS* software to extract the corresponding data.

3.5.4 Merging and Creating Variables

To obtain our final dataset, I merged the data on listings with the data on location premiums and created several new variables. I merged the two datasets on `ZSP` and `time_of_listing`, i.e. I matched every listing with the appropriate location premium based on its geographical location and the time of its listing. Finally, the following variables were created:

- `m2rent`: rent per square meter
- `cap_min`: individual rent cap in the pessimistic scenario
- `cap_max`: individual rent cap in the optimistic scenario

- `cap_real`: individual rent cap in the realistic scenario
- `over_min`: dummy which takes the value 1 if rent is above the rent cap in the pessimistic scenario
- `over_max`: like `over_min` but for the optimistic scenario
- `over_real`: like `over_min` but for the realistic scenario
- `size130`: dummy which takes the value 1 if apartment size is below $130m^2$

3.6 Empirical Strategy

The empirical strategy is threefold, testing the three hypotheses H$1-3$. The following sections lay out the strategies for each of the hypotheses and present the corresponding results.

3.6.1 Hypothesis H1

To test the first hypothesis, I compute the following descriptive statistics. Using data on apartment features and the corresponding surcharges and deductions, I compute apartment-specific rent caps. By comparing these computed rent caps to the listed rents, I observe whether a given listing complies with the rent regulation or not. According to hypothesis H1, the observed compliance rate should be close to zero. Since not all features that affect the apartment-specific rent cap are observable from the data, I perform this analysis for the three different scenarios described in Section 3.5.2.

In Figure 3.3, I plot a histogram of posted rents over rent caps. If this ratio is above 1, i.e. right of the red line, then the listing does not comply with the regulation. Figure 3.3 shows a distribution of these ratios for the realistic scenario while the pessimistic and optimistic scenarios are depicted in Figure C.8 in Appendix 3.B. As we can see, the majority of ratios lie to the right of the right line. In the realistic scenario, the share of non-compliant listings is 86%. In the optimistic and pessimistic scenarios, these shares are 76% and 96%, respectively.

There are three potential reasons for a strictly positive share of compliant listings. First, some landlords might simply be reluctant to violate the rent cap for ethical reasons. Second, for some apartments, the rent cap might not be binding because of their poor condition or undesirable location. Third, our data is subject to measurement error, and we do not observe all characteristics that determine the apartment-specific rent cap. In particular, as pointed out in Section 3.5.2, I do not observe whether the contract has

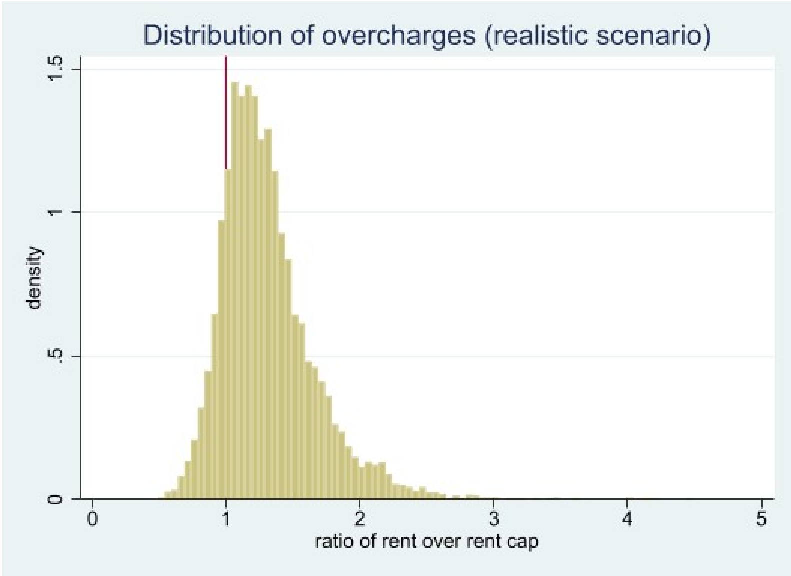


Figure 3.3: The distribution of rent-over-rent cap ratios for apartments that are subject to rent control. A significant majority of rental listings do not comply with the rent cap.

a fixed term which means that the estimated share of non-compliant listings is a lower bound.

Figure 3.4 shows the share of non-compliant listings in districts $2-23$ in the realistic scenario. The bars represent the shares in each respective district and the dashed vertical line shows the average over districts. Although we can see variation across districts, overcharging seems not to be driven by listings in certain districts.[25] Rather, it is a widespread phenomenon that occurs throughout the whole city. Overall, these findings support hypothesis H1: By and large, owners do not comply with the rent regulation.

3.6.2 Hypothesis H2

According to the second hypothesis, regulated apartments are more expensive than they would be without regulation. The applicability of the rent cap depends, among other things, on the size of the apartment. Only apartments between 30 and 130 m^2 are subject to the regulation. To test H2, I exploit the discontinuity at the upper threshold to estimate the local average treatment effect (LATE) using a regression discontinuity design. In the simplest version of it, I will estimate Equation (3.11) on a subset of observations including apartments within symmetric intervals around the threshold.

$$y_i = \alpha + \gamma_0 \texttt{size130}_i + \gamma_1(\texttt{size}_i - 130) + \gamma_2(\texttt{size}_i - 130) \times \texttt{size130}_i + \epsilon_i \qquad (3.11)$$

The outcome variable y_i on the left-hand side is apartment i's rent per square meter, $\texttt{size}_i$ is apartment i's floor area, and $\texttt{size130}_i$ is the treatment dummy, taking the value 1 if

[25]Proportion tests show that the shares in different districts differ significantly from each other.

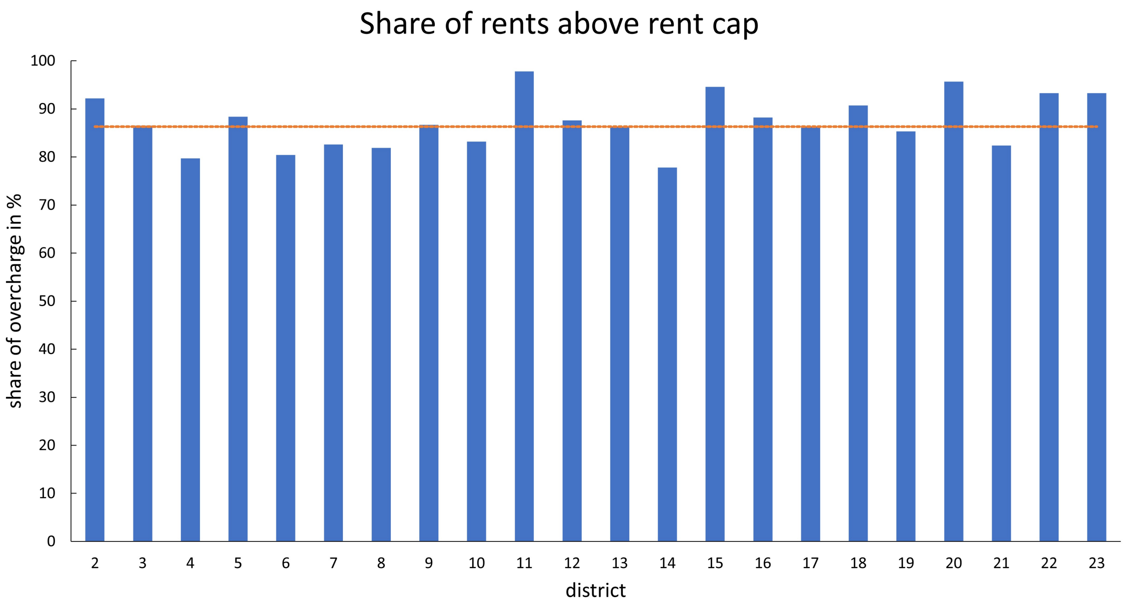


Figure 3.4: The blue bars show the shares of listings with a posted rent above the legal rent cap for districts $2-23$ in the realistic scenario. The dashed orange line shows the average over districts $2-23$.

apartment i is above the threshold of $130\,m^2$. Following Gelman and Imbens (2019), I also estimate a version including quadratic trends and a triangular kernel function. As we can see in Figure 3.6.2, this latter version is graphically appealing. Inspecting the estimated function and treatment effect, the quadratic trend and the (absence of a) treatment seem to fit the data well. I test alternative specifications, varying the order of the polynomial and the bandwidth and obtain similar estimates. All estimations are performed in *Stata* using the `rdrobust` package (Calonico et al., 2017).

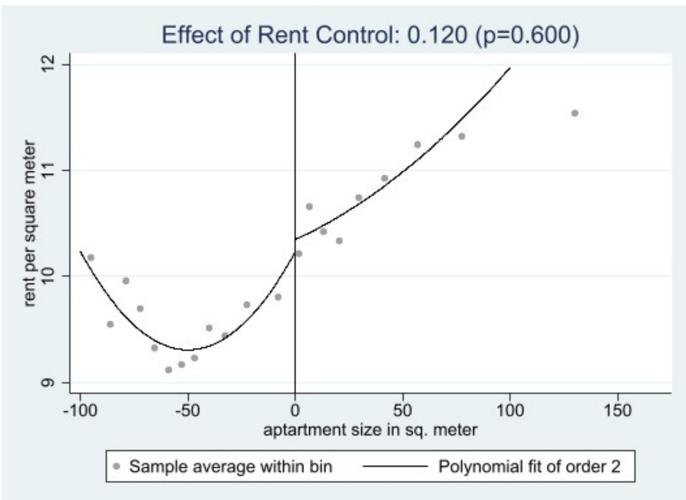


Figure 3.5: The plotted functions correspond to column (2) in Table 3.2. The data were binned such that each dot represents the mean of an equal number of observations.

It is important for the validity of the design that there is no manipulation of the running variable. Such manipulation could occur when landlords "cheat" and report a larger apartment size in order to make it above the threshold and discourage informed tenants from renting their apartment in hope of reducing the rent ex-post. To test for such manipulation, I perform a density test based on Cattaneo et al. (2020). The results are reported in Figure 3.6, which plots the apartment size distributions along with the estimated density functions and their 95%-confidencde intervals. The test rejects the null hypothesis of no discontinuity when performed on the full sample. It is important to note that the distribution of apartment sizes has peaks at multiples of 5 and even more pronounced peaks at multiples of 10. These peaks are most likely the result of rounding and can be seen very clearly in Figure C.9 in Appendix 3.B. Therefore, I expect the test to reject the null hypothesis too often. This is confirmed by the fact that the test rejects the null at 6 out of the 17 multiples of 10 in $[40, 200]$. Nevertheless, I choose to err on the side of caution and restrict the sample by excluding observations around the potential discontinuity. Repeating the test on this restricted sample, the null can no longer be rejected (see right panel of Figure 3.6).

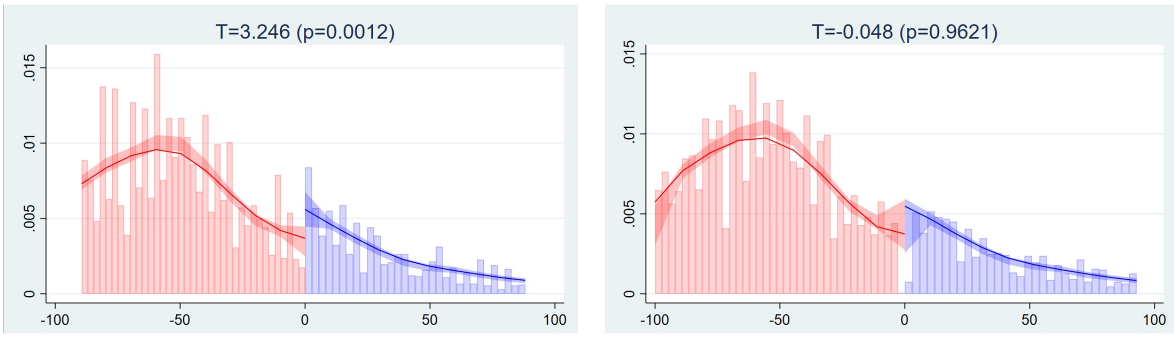


Figure 3.6: Density-based manipulation test. The estimated treatment effect for the full sample is positive and significant suggesting a discontinuity in the distribution of apartment sizes (left). The estimate for the donut hole sample $(128 - 132m^2)$ is insignificant (right).

Another central assumption behind the RDD is that at the threshold where the treatment probability jumps, all other observable or unobservable characteristics change smoothly. I test this assumption in Appendix 3.A, where I estimate the treatment effect on all observable characteristics and find no significant effects.

I report the estimates based on the full sample as well as the "donut hole" estimates in Table 3.2. Column (1) shows the estimate of the coefficient γ_0 in Equation (3.11). According to hypothesis $H2$, apartments that are subject to rent control have higher posted rents. However, all four specifications in Table 3.2 yield coefficient estimates that are not significantly different from 0.

At this point, an important caveat is due. We are not actually testing hypothesis $H2$ with our RDD. The reason is that $H2$ follows from the comparative statics of our model. It states that in a market with rent control, the rent is higher than it would be *in*

Table 3.2: Regression Discontinuity Design

Rent per m^2	(1)	(2)	(3)	(4)
Size130	0.181	0.120	0.353	0.140
	(.190)	(0.228)	(0.215)	(0.272)
Specification:				
linear trend	Yes		Yes	
bandwith [$\pm 50\, m^2$]	Yes		Yes	
quadratic trend		Yes		Yes
bandwith [$\pm 100\, m^2$]		Yes		Yes
full sample	Yes	Yes		
donut hole: $128 - 132\, m^2$			Yes	Yes
(effective) N	2923	5,754	2,769	5,600

Notes. The treatment variable Size130 indicates that an apartment is larger than $130\, m^2$.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

that same market without rent control. While the model does feature a regulated and an unregulated segment within the same market, the rent in the unregulated segment is exogenously fixed and serves the role of an outside option for tenants in the regulated market. Therefore, the results of our RDD should not be interpreted as a rejection of $H2$. This is an important caveat that will be addressed further in Section 3.7.

3.6.3 Hypothesis H3

The model in Section 3.4 predicts a negative relationship between the rent cap and posted rents. In this section, I investigate the hypothesis that an increase in rent cap leads to a decrease in posted rents. The empirical approach exploits the fact that the rent cap varies spatially across the city. I will first estimate a hedonic regression including the location premium as a hedonic variable. Finally, I will introduce a spatial discontinuity design.

Hedonic Price Regression

Before introducing the spatial discontinuity design, I first estimate the hedonic regression that is shown in Equation (3.12). It estimates the effect of the location premium on the per m^2 rent by including the location premium as a hedonic variable in an otherwise standard hedonic regression. The dependent variable y_i is the per square meter rent of apartment i. $\texttt{Premium}_i$ is the location premium that applies to apartment i, while the vector $\mathbf{x_i}$ contains all other hedonic variables. As before, $\mathbf{d_i}$ and $\mathbf{t}_i$ are the spatial and time dummies, respectively.

$$y_i = \beta \mathbf{x}_i + \gamma \text{Premium}_i + \delta \mathbf{d}_i + \tau \mathbf{t}_i + \epsilon_i \tag{3.12}$$

The coefficient of interest, γ, measures the effect of a 1€ increase in location premium on posted rents. According to hypothesis $H3$, an increase in the rent cap leads to a decrease in posted rents. Therefore, I expect a negative estimate for γ.

While the hedonic regression serves the purpose of giving us a first idea of the effect of rent caps, it is clear that it suffers from endogeneity. The problem is that location premiums are not allocated randomly and likely correlate with unobserved apartment features. If, for example, apartments in higher location premium zones have more (less) attractive unobserved features, our coefficient γ will overestimate (underestimate) the positive effect or underestimate (overestimate) the negative effect of an increase in the location premium. It could be that apartments in some parts of the city are more desirable due to their proximity to parks or other points of interest. If location premiums were chosen in order to reflect this desirability, γ will be biased.

To address this issue, I estimate Equation (3.12) with two different vectors of spatial control, one with district-fixed effects and one with ZSP-fixed effects. Unfortunately, including ZSP as spatial controls also controls for a large part of the variation in `Premium`, namely all variation between ZSP that is constant over time. Even though there is some variation in `Premium` over time as several ZSP change location premium zones, I am concerned that this residual variation is not sufficient to identify the effect of `Premium`.

Table 3.3 shows the estimation results for Equation (3.12). The only difference between columns (1) and (2) is the set of spatial control dummies. When controlling for district fixed effects, as in column (1), we obtain a significant and positive effect of the location premium of 0.66€. If rent control was complied with, the positive effect would be mechanical, and with less than perfect compliance the effect would then be smaller than 1. As we have seen in Figure 3.4, compliance is close to zero and can thus not explain the estimated positive effect. What might be driving the effect is that locations with higher premiums are more desirable for reasons that we are not able to control for and that this is true even within districts.

In column (2) of Table 3.3, I include ZSP fixed effects to control for any unobserved location-specific characteristics. Doing so controls for a large part of the variation in `premium`, namely that between ZSP, and leaves only the residual variation that comes from ZSP changing location premium zones to identify the effect of `premium`. As a result, the effect of the location premium disappears as we can see in column (2) of Table 3.3. This could either mean that our estimate in column (1) was driven entirely by the bias or

Table 3.3: Hedonic regression

Rent per m^2	(1)	(2)
Premium	0.660***	-0.071
	(0.062)	(0.096)
Controls:		
district FE	Yes	
ZSP FE		Yes
R^2	0.261	0.570
N	4,365	4,365

Notes. OLS estimation of Eq. (3.12). A fulll list of controls and their estimated effects can be found in Appendix 3.B.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

that there is not enough residual variation left after controlling for ZSP to identify the effect of interest.

Spatial Discontinuity Design

As mentioned above, which ZSP is allocated to which location premium is not random. In particular, areas with higher location premiums are probably more desirable overall and apartments in such regions are likely to be of higher quality in unobserved ways. In the spirit of RDD, I use a spatial discontinuity design to estimate the effect of an increase in location premium at the border between two location premium zones. To this end, I estimate Equation (3.12) restricting the sample to observations that are located at a border between two premium zones. More precisely, I include only apartments in ZSP that share a border with a ZSP where a different location premium applies. Coefficient γ thus represents the average treatment effect (ATE) of a 1€ increase in the rent cap at the border between different location premium regions.

The identifying assumption behind my spatial discontinuity design is that, unlike the location premium, other - and in particular unobserved - apartment characteristics do not change discontinuously when crossing a border. Because such borders tend to be rather long it is likely that apartment characteristics change *along* a border. To control for such changes, I include a set of cross-regional neighborhood dummies that were constructed using *QGIS*. A neighborhood consists of ZSPs that share a substantial border and importantly includes ZSP on both sides of a border.

To allow for the possibility that a region border that simultaneously separates two districts is more substantial than one that does not, I estimate two specifications, one

Table 3.4: Spatial Discontinuity Design

Rent per m^2	(1)	(2)
Premium	0.257^{**}	0.212^{*}
	(0.104)	(0.116)
Controls:		
district FE		Yes
R^2	0.526	0.533
N	3,172	3,172

Notes. OLS estimation of Eq. (3.12) on a restricted sample that includes only apartments on borders between different location premium zones. Full list of controls and their estimated effects can be found in Appendix 3.B.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

with and one without district-fixed effects.

The results are shown in Table 3.4 and the coefficient of interest is again that of `Premium`, γ. Column (1) of Table 3.4 shows a positive treatment effect at premium region borders. Apartments on the side of the border with the higher rent premium are, on average, 0.26€ $/m^2$ more expensive for every 1€ difference in the location premium. As we can see in column (2), this effect is robust to the inclusion of district-fixed effects even though it shrinks in magnitude (0.21€ $/m^2$) and significance. The results are also robust to different polynomials in `size`.

As pointed out earlier in Section 3.6.2, the predictions of the theoretical model come from its comparative statics. Therefore, Hypothesis $H2$ predicts that *ceteris paribus* a market with a lower rent cap will have higher posted rents. Instead, we are looking at the effect of differential rent caps within the same market. I will address this mismatch in further work, as explained in the following section.

3.7 Further Steps

In the future, I will use new data on rental listings. These new data have two important advantages. The first one is their amount and quality. Because the new data set is generated by a software used by real estate agents to place listings on different platforms rather than extracting them from these platforms, duplicates will not be of major concern. The second advantage is that the new data has apartment locations on the address level instead of the ZSP level. This not only allows me to restrict the sample to a smaller distance around location premium borders but also to employ a new specification altogether. With precise location, I can use a parametric spatial discontinuity design by including distance to the border as a running variable. The increased number of observations will also allow me to restrict the sample to those border regions that are not district borders. Ultimately, with enough data, I will be able to restrict the sample to only include one border between two regions at a time. That is, I will estimate a border-specific treatment effect instead of an average treatment effect across all borders.

The most important aspect of future work, however, will be addressing the mismatch between the theoretical analysis and the empirical strategy. While the simple model presented in this chapter serves the purpose of shining a first light on imperfectly enforced rent control, its predictions do not permit a straightforward econometric investigation. I will address this problem from the theoretical side by adapting the model such that multiple sectors with different rent caps co-exist in the same market. Preliminary work suggests that the main result, that lax rent control leads to an increase in posted rents, will continue to hold. The key to this result is the following: Lax rent control reduces supply because landlords expect ex-postnforcement with positive probability. However, the fact that it is not complied with a priori means that the rent level can adjust to equilibrate demand and supply. This necessarily leads to an increase in posted rents.

3.8 Conclusion

Rent control is not only a controversial topic among economists, but it is also a subject of heated political discussions. The effectiveness of the Austrian rent regulation has long been questioned. This study showed that these concerns are justified. I exploited two institutional details of the regulation to utilize a novel identification strategy. First, I established that a large majority of listings do not comply with the rent regulation.

I then made use of the sharp threshold of $130\,m^2$ above which rent caps no longer apply, and utilized an RD design to estimate the local average treatment effect of rent control. My findings show that for apartments close to the threshold, the posted rent is independent of the side of the threshold they are on. Second, using the variation of the rent cap between geographical regions and the fact that borders between those regions are sharp, I estimated that the average effect of a change in the rent cap is at best moderate. Thus, the findings presented in this paper suggest that the Austrian rent regulation is indeed not very effective.

3.A Balance Checks

In this appendix, I want to assess the validity of our empirical approach. Recall that the identifying assumption behind our regression and spatial discontinuity designs is that apartment characteristics do not change in a discontinuous fashion when the relevant threshold is crossed. In our RD design, this threshold corresponds to an apartment size of $130\,m^2$, while in our SD design, the threshold is the border between two regions of different location premiums.

Before turning to balance checks for these specifications, I first look at the balance in the hedonic regression specification from Section 3.6.3. As mentioned in this section, endogeneity is likely to bias the estimates because location premiums are not allocated randomly. To verify this claim and to provide a point of reference for further balance checks, I estimate the linear probability model in Equation C.13.

$$y_i = \alpha + \beta \text{Premium}_i + \epsilon_i \tag{C.13}$$

I estimate Equation C.13 for each of the control variables separately. The results are shown in column (3) of Table C.5. Note that stars indicate significance on an individual level while plusses indicate significance after the Bonferroni correction for multiple hypothesis tests. Coefficients that are statistically significant after the Bonferroni correction are in **bold**. As we can see, the location premium has high predictive power for the majority of our covariates. This is not very surprising and the reason why I resort to specification S3 to overcome this issue.

Turning to the spatial discontinuity design, I investigate to what extent it is plagued by the same endogeneity problem. To this end, I estimate the following linear probability model in Equation C.14, where y again corresponds to a covariate and the variable `expensive` is a treatment dummy that takes the value 1 whenever the apartment is on the more expensive side of a border. I estimate Equation C.14 first without spatial controls, i.e. with $\mathbf{d}_i = \vec{0}$ for sake of better comparison with column (3). The results are shown in column (4) of Table C.5 and we can see that the endogeneity is still present albeit to a lesser extent with only 5 out of the 31 covariates being significantly different from 0.

$$y_i = \alpha + \beta \text{expensive}_i + \delta \mathbf{d}_i + \epsilon_i \tag{C.14}$$

However, we are interested in whether apartments on opposite sides of a border *within a neighborhood* differ systematically. I, therefore, estimate Equation C.14 with neighborhood fixed effects and present the findings in column (5). As a result, the explanatory power

Table C.5: Balance checks for specifications S1-3

		S1		S2	S3	
Controls:		(1)	(2)	(3)	(4)	(5)
area		-	-	**7.320*****	**8.655*****	**7.007*****
condition	0	-0.018	-0.012	**0.018*****	**0.047*****	0.035
	1	-0.001	0.000	-0.000	-0.000	-0.000
	2	-0.010	-0.010	-0.005	-0.004	0.009
	3	-0.016	0.029	**0.060*****	**−0.063*****	-0.016
	4	0.045	0.051	**0.047*****	0.018	-0.028
elevator		0.022	0.020	**0.136*****	**0.066*****	0.041
basement		-0.008	-0.054	-0.021	-0.037	-0.053
garage		0.032	0.040	**0.016*****	-0.012	-0.020
no. of balconies		0.041	0.034	**0.060*****	**0.067*****	0.026
no. of terraces		-0.002	-0.044	0.016	0.008	0.045
no. of loggias		0.030	0.016	0.003	0.001	0.002
no. of bathrooms		-0.003	0.058	**0.020*****	-0.007	-0.002
no. of toilets		0.086	-0.103	**0.053*****	0.013	-0.024
year	2010	-0.079	-0.062	**−0.067*****	0.003	0.040
	2011	0.023	0.045	**−0.071*****	0.000	-0.028
	2012	0.006	0.009	**−0.044*****	-0.015	0.021
	2013	0.048	-0.038	**−0.032*****	0.008	-0.011
	2014	0.055	0.060	**0.025*****	0.003	-0.033
	2015	-0.063	-0.074	**0.075*****	0.014	0.008
	2016	0.032	0.017	**0.056*****	-0.001	0.002
	2017	-0.022	-0.034	**0.057*****	-0.006	0.001
central heating		0.066	0.021	0.017	0.013	-0.052
floor	0	0.010	0.011	-0.009	-0.012	-0.025
	1	0.020	0.009	0.015	0.021	0.043
	2	-0.008	-0.002	-0.007	-0.026	0.012
	3	-0.015	-0.011	-0.008	-0.010	0.010
	4	0.007	0.029	0.006	-0.005	-0.004
	5	-0.011	-0.029	0.003	-0.003	0.001
	6	-0.001	-0.005	-0.000	-0.002	0.003
	7	-0.002	-0.002	0.000	0.000	0.004
no. of observations		5928	5774	4,365	3,182	3,182

Notes. Balance checks for our research designs.
Bonferroni-corrected significance levels: * $p < 0.0036$, ** $p < 0.0018$, *** $p < 0.0004$

of the treatment variable `expensive` decreases further. However, it still explains the variation of `size` to a significant extent. I conclude that the treatment and control groups a sufficiently balanced to support our SD design and that controlling for apartment size is important to reduce the potential bias further.

Finally, we turn to the regression discontinuity design. In order to assess the balance of apartments on either side of the threshold, I run an RDD for each of the covariates. I use the quadratic specifications corresponding to columns (2) and (4) of Table 3.2, i.e. column (1) shows the RD estimates for the full sample and column (2) for the donut hole sample. While some of the estimates are significant at the individual level, none of them are after correcting for multiple hypothesis tests. I am thus confident that my sample is reasonably balanced for the RD design to be valid.

3.B Tables and Figures

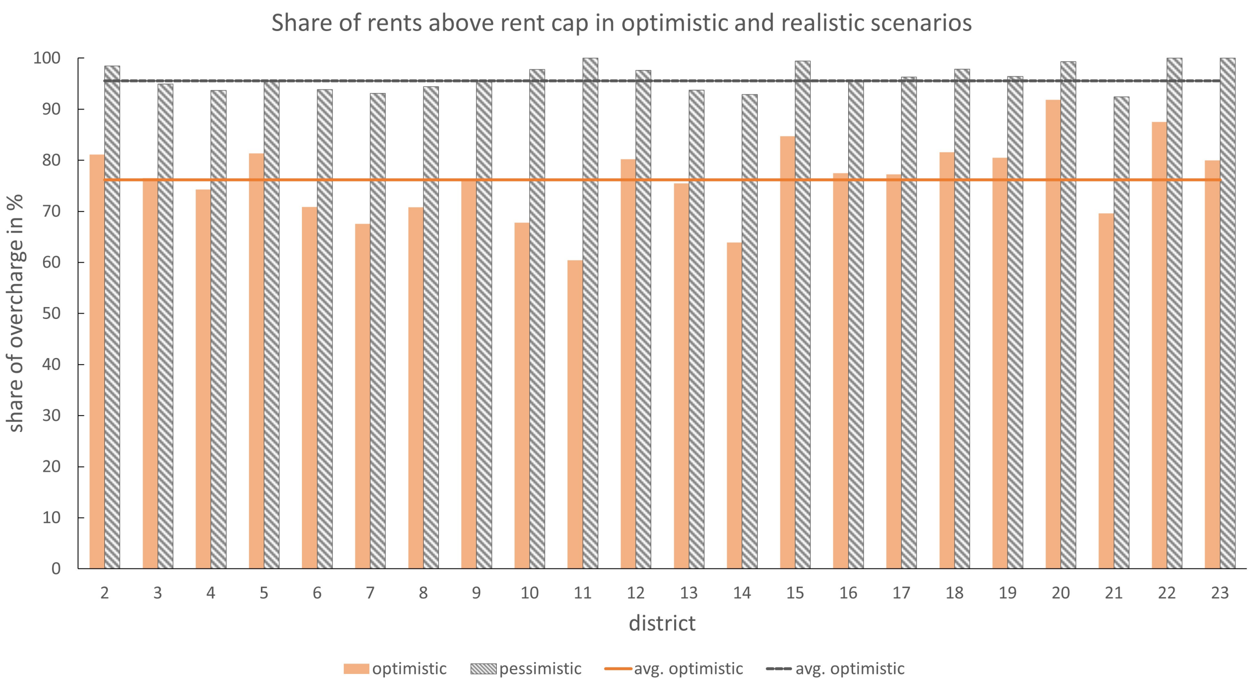


Figure C.7: Orange solid bars show the shares of overcharging listings in districts $2-23$ for the optimistic scenario. The solid orange line shows the average over districts $2-23$ for the optimistic scenario. The shaded bars and dashed line show the pessimistic scenario.

Table C.6: Coefficients of control variables in Tables 3.3 - 3.2

		S1		S2		S3	
Controls:		(1)	(2)	(1)	(2)	(1)	(2)
area		0.023*	0.023*	-0.058***	-0.060***	-0.061***	-0.061***
area^2		0	0	-0.0002***	0.0002***	0.0002***	0.0002***
condition	1	-2.125***	-1.614***	-2.261***	-2.254***	-2.569***	-2.508***
	2	-2.248***	-2.080***	-2.204***	-1.875***	-1.592***	-1.581***
	3	-0.973***	-0.876***	-0.848***	-0.718***	-0.763***	-0.757***
	4	0.236***	0.197**	0.314***	0.261**	0.333***	0.331***
basement		-0.117**	-0.062***	-0.255***	-0.135**	-0.148*	-0.161**
elevator		0.507***	0.578***	0.502***	0.544***	0.402***	0.378***
# of balconies		0.239***	0.228***	0.464***	0.437***	0.538***	0.556***
# of loggias		0.023	0.010	0.306***	0.625**	0.655**	0.723***
# of terraces		0.845***	1.139***	0.687***	1.377***	2.104***	2.115***
# of bathrooms		0.524***	0.475***	0.582***	0.498***	0.328*	0.337***
# of toilets		0.749***	0.587***	0.904***	0.653***	0.790***	0.787***
year	2011	0.285***	0.299***	0.123	0.202**	0.108	0.100
	2012	0.268***	0.345***	0.258**	0.352 ***	0.162	0.141
	2013	0.473***	0.574***	0.512***	0.656***	0.502***	0.495***
	2014	0.208**	0.614***	0.260**	0.701***	0.382***	0.420***
	2015	0.308***	0.749***	0.384***	0.836***	0.381***	0.441***
	2016	0.152	0.631***	0.193	0.702***	0.323**	0.375**
	2017	-0.044	0.709***	0.035	0.826***	0.374**	0.473***
central heating		0.162**	0.052	0.220***	0.005	0.069	0.066
floor	0	-0.028	-0.297*	-0.001	-0.247	-0.222	-0.235
	1	-0.109	-0.223***	0.022	-0.115	-0.005	0.016
	2	-0.331***	-0.240***	-0.239***	-0.124	-0.011	0.023
	3	-0.091	-0.058	-0.052	0.010	0.178	0.207*
	4	0.102	0.147	0.144	0.163	0.206	0.223
	5	1.918***	1.445***	2.463***	1.846***	1.669***	1.637
	6	-0.422	-0.251	-0.173	0.060	-0.100	-0.103
	7	3.188***	2.458***	3.766***	2.634*	1.048**	1.042**

Notes. Spatial control variables are omitted. Column numbers correspond to those in Tables 3.3 - 3.2
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

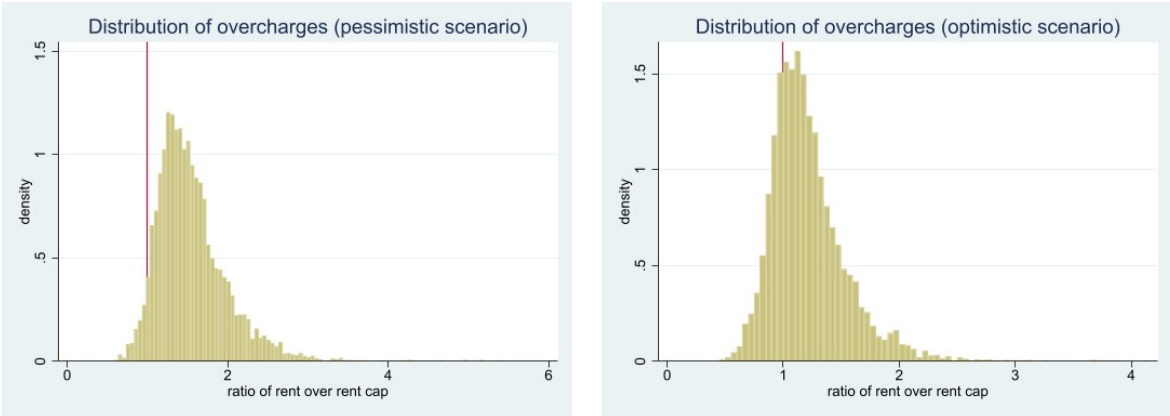


Figure C.8:

The distribution of rent-over-rent cap ratios for apartments that are subject to rent control. Left panel for the pessimistic scenario. Right panel for the optimistic scenario.

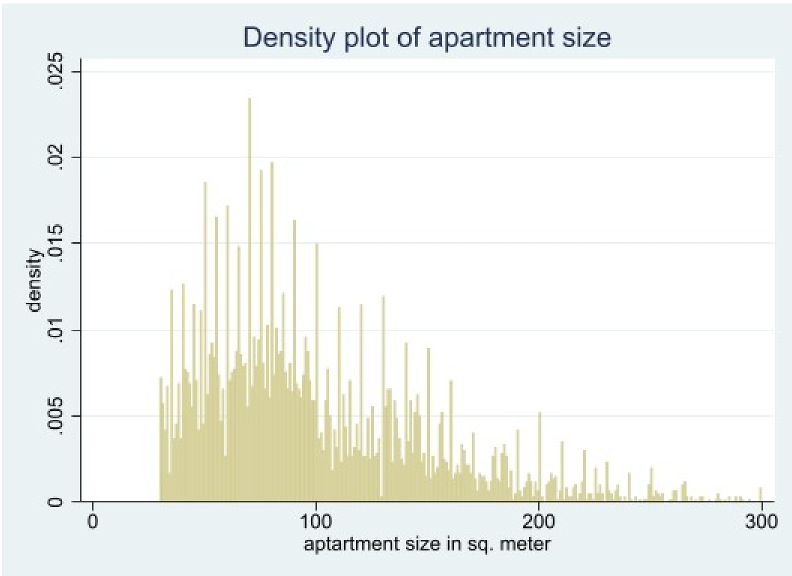


Figure C.9: The peaks correspond to multiples of 5.

Bibliography

Aguirregabiria, Victor, "Identification of firms' beliefs in structural models of market competition," *Canadian Journal of Economics*, February 2021, *54* (1), 5–33.

Alston, Richard M, James R Kearl, and Michael B Vaughan, "Is There a Consensus among Economists in the 1990's?," *The American Economic Review*, 1992, *82* (2), 203–209.

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

Aoyagi, Masaki, Guillaume R. Fréchette, and Sevgi Yuksel, "Beliefs in Repeated Games," *Working paper*, May 2022.

Arnott, Richard, "Time for revisionism on rent control?," *Journal of economic perspectives*, 1995, *9* (1), 99–120.

Ashenfelter, Orley and Robert S Smith, "Compliance with the minimum wage law," *Journal of Political Economy*, 1979, *87* (2), 333–350.

Au, Pak Hung and Keiichi Kawai, "Competitive information disclosure by multiple senders," *Games and Economic Behavior*, jan 2020, *119*, 56–78.

_ **and** _ , "Competitive disclosure of correlated information," *Economic Theory*, jan 2021, *72* (3), 767–799.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy, "Incentives in Experiments: A Theoretical Analysis," *Journal of Political Economy*, aug 2018, *126* (4), 1472–1503.

Basu, Arnab K, Nancy H Chau, and Ravi Kanbur, "Turning a blind eye: Costly enforcement, credible commitment and minimum wage laws," *The Economic Journal*, 2010, *120* (543), 244–269.

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).

Baumgartner, Josef, "Die Mietpreisentwicklung in Österreich," Technical Report July 2013.

Becker, Gordon M., Morris H. Degroot, and Jacob Marschak, "Measuring utility by a single-response sequential method," *Behavioral Science*, 1964, *9* (3), 226–232.

Bernheim, B. Douglas, Stefano DellaVigna, and David Laibson, eds, "Errors in Probabilistic Reasoning and Judgment Biases," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics*, Vol. 2, Elsevier, 2019, pp. 69–186.

Board, Simon and Jay Lu, "Competitive Information Disclosure in Search Markets," *Journal of Political Economy*, oct 2018, *126* (5), 1965–2010.

Bourjade, Sylvain and Bruno Jullien, "The roles of reputation and transparency on the behavior of biased experts," *The RAND Journal of Economics*, 2011, *42* (3), 575–594.

Breidenbach, Philipp, Lea Eilers, and Jan Fries, "Temporal dynamics of rent regulations–The case of the German rent control," *Regional Science and Urban Economics*, 2022, *92*, 103737.

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

Cai, Hongbin and Joseph Tao-Yi Wang, "Overcommunication in strategic information transmission games," *Games and Economic Behavior*, 2006, *56* (1), 7–36.

Calonico, Sebastian, Matias D Cattaneo, Max H Farrell, and Rocio Titiunik, "rdrobust: Software for regression-discontinuity designs," *The Stata Journal*, 2017, *17* (2), 372–404.

Carpenter, Jeffrey, Emiliano Huet-Vaughn, Peter Hans Matthews, Andrea Robbett, Dustin Beckett, and Julian Jamison, "Choice Architecture to Improve Financial Decision Making," *The Review of Economics and Statistics*, mar 2021, *103* (1), 102–118.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma, "Simple local polynomial density estimators," *Journal of the American Statistical Association*, 2020, *115* (531), 1449–1455.

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.

Charness, Gary, Uri Gneezy, and Michael A. Kuhn, "Experimental methods: Between-subject and within-subject design," *Journal of Economic Behavior & Organization*, jan 2012, *81* (1), 1–8.

Chen, Daniel L., Martin Schonger, and Chris Wickens, "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 2016, *9*, 88–97.

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

Chevalier, Judith A and Dina Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, 2006, *43* (3), 345–354.

Costa-Gomes, Miguela A. and Georg Weizsäcker, "Stated Beliefs and Play in Normal-Form Games," *Review of Economic Studies*, July 2008, *75* (3), 729–762.

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.

Danziger, Leif, "Endogenous monopsony and the perverse effect of the minimum wage in small firms," *Labour Economics*, 2010, *17* (1), 224–229.

Deversi, Marvin, Alessandro Ispano, and Peter Schwardmann, "Spin doctors: An experiment on vague disclosure," *European Economic Review*, 2021, *139*, 103872.

Edwards, Ward, Lawrence D. Phillips, William L. Hays, and Barbara Goodman, "Probabilistic Information Processing Systems: Design and Evaluation," *IEEE Transactions on Systems Science and Cybernetics*, September 1968, *4* (3), 248–265.

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

Feistritzer, Gert, "AK-Wien Studie Mietenbelastung," Technical Report February 2010.

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.

Frederick, Shane, “Cognitive Reflection and Decision Making,” *Journal of Economic Perspectives*, nov 2005, *19* (4), 25–42.

Fulghieri, Paolo, Günter Strobl, and Han Xia, “The Economics of Solicited and Unsolicited Credit Ratings,” *Review of Financial Studies*, February 2014, *27* (2), 484–518.

Gelman, Andrew and Guido Imbens, “Why high-order polynomials should not be used in regression discontinuity designs,” *Journal of Business & Economic Statistics*, 2019, *37* (3), 447–456.

Giacomelli, Silvia and Carlo Menon, “Does weak contract enforcement affect firm size? Evidence from the neighbour’s court,” *Journal of Economic Geography*, 2017, *17* (6), 1251–1282.

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,” *The Economic Journal*, 2021, *131* (636), 1772–1787.

Greenwald, Anthony G., “Within-subjects designs: To use or not to use?,” in “Methodological issues & strategies in clinical research.,” American Psychological Association, 1992, pp. 157–167.

Häckner, Jonas and Sten Nyberg, “Rent Control and Prices of Owner-occupied Housing,” *Scandinavian Journal of Economics*, 2000, *102* (2), 311–324.

Hahn, Anja M, Konstantin A Kholodilin, Sofie R Waltl, and Marco Fongoni, “Forward to the past: Short-term effects of the rent freeze in Berlin,” *Management Science, forthcoming*, 2023.

Harrison, Glenn, Jimmy Martínez-Correay, Karlijn Morsinkz, Jia Min Ng, and Todd Swarthout, "Compound Risk and the Welfare Consequencesof Insurance," *Working Paper*, December 2020.

_ , **Karlijn Morsink, and Mark Schneider**, "Do No Harm? The Welfare Consequences of Behavioural Interventions," *Working Paper*, July 2020.

Hayek, Friedrich A Von, "The use of knowledge in society," *American economic review*, 1945, *35* (4), 519–530.

He, Sherry, Brett Hollenbeck, and Davide Proserpio, "Exploiting Social Media for Fake Reviews: Evidence from Amazon and Facebook," *ACM SIGecom Exchanges*, November 2021, *19* (2), 68–74.

_ , _ , **and** _ , "The market for fake reviews," *Marketing Science*, 2022, *41* (5), 896–921.

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.

Ho, Lok Sang, "Rent control: its rationale and effects," *Urban Studies*, 1992, *29* (7), 1183–1189.

Janssen, Maarten CW and Santanu Roy, "Regulating product communication," *American Economic Journal: Microeconomics*, 2022, *14* (1), 245–83.

Jendraszek, Patricia Anne, "Misconceptions of Probability Among Future Teachers of Mathematics." PhD dissertation, Columbia University 2008.

Jenkins, Blair, "Rent control: Do economists agree?," *Econ journal watch*, 2009, *6* (1), 73.

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.

Kawagoe, Toshiji and Hirokazu Takizawa, "Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information," *Games and Economic Behavior*, 2009, *66* (1), 238–255.

Kettunen, Hanna and Hannu Ruonavaara, "Rent regulation in 21st century Europe. Comparative perspectives," *Housing Studies*, 2021, *36* (9), 1446–1468.

Kholodilin, Konstantin A, "Rent control effects through the lens of empirical research," Technical Report 2022.

Knapp, Boris, "Biased Experts and Naive Consumers," *Vienna Economics Papers Working Paper*, 2020.

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

Kutty, Nandinee K, "The impact of rent control on housing maintenance: A dynamic analysis incorporating European and North American rent regulations," *Housing Studies*, 1996, *11* (1), 69–88.

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

Lemos, Sara, "Minimum wage effects in a developing country," *Labour Economics*, 2009, *16* (2), 224–237.

Lewis, Gregory and Georgios Zervas, "The supply and demand effects of review platforms," *Available at SSRN 3468278*, 2019.

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

Martin, Simon and Sandro Shelegia, "Underpromise and overdeliver?-Online product reviews and firm pricing," *International Journal of Industrial Organization*, 2021, *79*, 102775.

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

Mense, Andreas, Claus Michelsen, and Konstantin Kholodilin, "Empirics on the causal effects of rent control in Germany," *Beiträge zur Jahrestagung des Vereins für Socialpolitik 2018: Digitale Wirtschaft - Session: Inequality III,*, 2018, *No. E04-V3.*

Moon, Choon-Geol and Janet G Stotsky, "The effect of rent control on housing quality change: a longitudinal analysis," *Journal of Political Economy*, 1993, *101* (6), 1114–1148.

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.

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

Olsen, Edgar O, "What do economists know about the effect of rent control on housing maintenance?," *The journal of real estate finance and economics*, 1988, *1*, 295–307.

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

_ **and Peter Norman Sørensen**, "Reputational Cheap Talk," *The RAND Journal of Economics*, 2006, *37* (1), 155–175.

Phillips, Lawrence D. and Ward Edwards, "Conservatism in a Simple Probability Inference Task," *Journal of Experimental Psychology*, 1966, *72* (3), 346–354.

Pichlmair, Michael, *Miete, Lage, Preisdiktat: Strukturelle Effekte der Lageregulierung im mietrechtlich geschützten Wiener Wohnmarkt*, Peter Lang International Academic Publishers, 2013.

Raven, J. C., J. Raven, and J. H. Court, *A manual for Raven's Progressive Matrices and Vocabulary Scales.*, London: H. K. Lewis, 1998.

Rey-Biel, Pedro, "Equilibrium play and best response to (stated) beliefs in normal form games," *Games and Economic Behavior*, mar 2009, *65* (2), 572–585.

Rosifka, Walter and René Postler, "Die Praxis des Richtwertmietzinssystems," Technical Report December 2010.

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.

Schlag, Karl and James Tremewan, "Simple belief elicitation: An experimental evaluation," *Journal of Risk and Uncertainty*, apr 2021, *62* (2), 137–155.

Schneider, Ulrich C., "Identifying and Estimating Beliefs from Choice Data - An Application to Female Labor Supply," *Working paper*, February 2019.

Sims, David P, "Out of control: What can we learn from the end of Massachusetts rent control?," *Journal of Urban Economics*, 2007, *61* (1), 129–151.

Smirnov, Aleksei and Egor Starkov, "Bad news turned good: reversal under censorship," *American Economic Journal: Microeconomics*, 2022, *14* (2), 506–560.

Sobel, Joel, "A Theory of Credibility," *The Review of Economic Studies*, oct 1985, *52* (4), 557.

Stigler, George J, "The economics of information," *Journal of political economy*, 1961, *69* (3), 213–225.

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

Abstract

This dissertation provides a comprehensive exploration of information provision and consumer protection. It comprises three distinct chapters.

The first chapter studies the interplay between naive consumers and fake reviews within a theoretical model. Through an examination of awareness policies and their impact on both sophisticated and naive consumers, the chapter establishes that these policies inevitably trade off the welfare of one group against the other, but ultimately increase aggregate consumer surplus. The chapter also sheds light on the incentives of real reviewers in such markets and establishes that if they write reviews strategically, they are not always truthful.

The second chapter, a collaborative effort with Dominik Stelzeneder, investigates the effects of informational treatments within an asymmetric information game using an online experiment. The chapter illustrates how the presence of fake reviews is not only a piece of information that consumers can be misinformed about, it is also something that complicates consumers' information extraction process. It demonstrates that consumers' decisions are significantly impaired when they are unable to correctly compute posterior probabilities, and highlights the importance of combining awareness policies with instruments that help consumers incorporate the information correctly.

The third chapter focuses on the effects of an imperfectly enforced rent control policy, using the Austrian rent regulation as an example. The chapter introduces a theoretical model that highlights how such a regulation can lead to an unintended increase in rents. The chapter provides partial support for these hypotheses using data from rental listings in Vienna and regression and spatial discontinuity designs.

Overall, this dissertation thoroughly analyses the complex relationship between information provision and consumer protection. The findings have significant implications for policymakers, as they provide valuable insights into the challenges of regulating online platforms and enforcing rent control policies.

Zusammenfassung

Die vorliegende Dissertation bietet eine Erkundung von Informationsbereitstellung im Zusammenhang mit Verbraucherschutz. Sie besteht aus drei eigenständigen Aufsätzen.

Der erste Aufsatz untersucht das Zusammenspiel von naiven Verbrauchern und gefälschten Bewertungen innerhalb eines theoretischen Modells. Durch eine Untersuchung von Aufklärungsinitiativen und ihrem Einfluss auf sophistizierte und naive Verbraucher, zeigt das Kapitel auf, dass diese Politiken zwangsläufig das Wohlergehen einer Gruppe gegenüber das der anderen abtauschen, aber letztendlich die Konsumentenrente erhöhen. Das Kapitel beleuchtet auch die Anreize realer Bewertungen in solchen Märkten und zeigt auf, dass wenn sie strategisch geschrieben werden, sie nicht immer wahrheitsgemäß sind.

Der zweite Aufsatz, eine Zusammenrbeit mit Dominik Stelzeneder, untersucht die Auswirkungen von Informationsbereitstellung mithilfe eines Online-Experiments. Das Kapitel verdeutlicht, wie die Präsenz von gefälschten Bewertungen nicht nur Information ist, welche Verbraucher oft fehlt, sondern auch etwas, das den Informationsgewinnungsprozess der Verbraucher erschwert. Es zeigt auf, dass die Entscheidungen der Verbraucher erheblich beeinträchtigt werden, wenn sie nicht in der Lage sind, bedingte Wahrscheinlichkeiten korrekt zu berechnen. Das Kapitel verdeutlicht die Wichtigkeit Aufklärungspolitiken mit Instrumenten zu kombinieren, die den Verbrauchern helfen, die Informationen korrekt zu intepretieren.

Der dritte Aufsatz befasst sich mit den Auswirkungen einer lückenhaft durchgesetzten Mietpreisregulierung. Der empirische Teil untersucht die Auswirkungen des österreichischen Mietrechtsgesetz auf den Wiener Mietmarkt. Der Aufsatz stellt ein theoretisches Modell vor, das aufzeigt, wie eine solche Regulierung zu einer unbeabsichtigten Erhöhung der Mieten führen kann. Mit Hilfe von Daten aus Mietangeboten in Wien liefert das Kapitel teilweise Unterstützung für diese Hypothesen.

Insgesamt liefert diese Dissertation eine umfassende Analyse der komplexen Beziehung zwischen Informationsbereitstellung und Verbraucherschutz. Die Ergebnisse sind von unmittelbarer Bedeutung für die Entscheidungen von Politikern, da sie wertvolle Einblicke in die Herausforderungen bei der Regulierung von Online-Plattformen und der Durchsetzung von Mietpreisregulierungen bieten.