When Market Unravelling Fails and Mandatory Disclosure Backfires: Persuasion Games with Labelling and Costly Information Acquisition

Ennio Bilancini[*] Leonardo Boncinelli[†]

January 12, 2021

Abstract

In this paper we develop a variant of the persuasion game by Milgrom and Roberts (1986) to study the emergence and the desirability of product labelling when buyers can acquire information on the quality of the product by paying a cost. Labelling is modeled as the (verifiable) public disclosure of an otherwise unobservable trait of the seller that is correlated with the quality of the product. Our main finding is that market unravelling can fail, in which case imposing mandatory disclosure can backfire. When the joint distribution of seller's qualities and traits is exogenous, if market unravelling fails and mandatory labelling is imposed, then profits decrease for high quality sellers and, if the label is sufficiently informative, buyers are better off and profits increase for low quality sellers as well. When, instead, the joint distribution of qualities and traits is endogenous, mandatory labelling fails to yield an increase in average quality and buyers' utility, while cost inefficiencies arise.

JEL classification code: D82, D83, L15.

Keywords: certification, cognitive resources, intuitive/reflective, mandatory disclosure, unraveling.

[*] IMT School of Advanced Studies, Piazza San Francesco 19, 55100, Lucca, Italy. Tel.: +39 0583 4326 737, email: ennio.bilancini@imtlucca.it.

[†] Department of Economics and Management, University of Florence, Via delle Pandette 9, 50127 Florence, Italy. Tel.: +39 055 2759578, email: leonardo.boncinelli@unifi.it.

1 Introduction

1.1 Motivation and setting

Mandatory disclosure of product information is often considered desirable, especially in consumer good markets (Dranove and Jin, 2010). An important justification for this is that asymmetric information about product quality and misalignment of incentives between sellers and buyers are often the rule, with the result that sellers withhold product information that, if instead were disclosed, would be beneficial to buyers.

However, as pointed out by marketing and psychological research (see, e.g., Loewenstein et al., 2014, and references therein), buyers' cognitive limitations such as scarce attention or costly elaboration of information can significantly impair, or even reverse, the intended effects of mandatory disclosure. More in general, in many cases buyers have the option to acquire information on the quality of the product by paying a cost – cognitive, material, or both – and this possibility to retrieve information on their own can be a substitute for disclosure, deeply altering the effects of mandatory disclosure. Also, it is typically not feasible for sellers to disclose exact information on quality.[1] At the same time, it is often feasible to disclose and certify some product trait that correlates with product quality, although this might require to incur a disclosure/certification cost.

In particular, there are situations where buyers can acquire information on product quality that are more precise than the information that can be disclosed by sellers through certification. One such case is when quality is subjective, depending on private characteristics of buyers, who are heterogeneous along these characteristics (e.g., preferences or contexts).[2] Another case is when quality is objective but too complex to be technically certifiable by sellers in a precise way, and some buyers have developed an expertise that allows them to infer quality with good precision if they exert effort.[3]

In either case the assessment of quality is often a costly task for a buyer, in that it requires a careful inspection of the product under consideration, and possibly some elaboration of the gathered information as well. Further, while buyers may acquire information on quality,

[1] Brécard (2017) studies the functioning of green markets, assuming that eco-labels distinguishing firms of different environmental quality are misperceived by consumers, who are hence unable to exploit such information.

[2] Guo and Zhang (2012) formalize costly consumer introspection for the determination of willingness to pay. Guo (2016) provides a framework for the contextual formation of preferences.

[3] An expert's assessment cannot be used by sellers to certify quality precisely because the assessment has a qualitative and hardly verifiable nature (Chakraborty and Harbaugh, 2010) and because of competition among experts (Schmidbauer, 2017).

they cannot acquire a seller's trait, since it is difficult to verify or certify it for a single buyer (e.g., country of origin, adherence to production standard, origin of raw materials, etc.).

In this paper we develop a variant of the persuasion game by Milgrom and Roberts (1986) with the aim of capturing the situations described above. In our model, a seller can be of high or low quality and has one of two traits that correlate differently with quality.[4] Initially, both quality and trait are private information of the seller. Before buyers decide whether to acquire information on quality, the seller can disclose and certify his own trait, incurring a disclosure cost. Such disclosure is called "labelling" in order to distinguish it from the direct disclosure of quality.[5] In this setup we study when voluntary labelling by sellers fails – i.e., there is not market unravelling – and, in such cases, to what extent mandatory labelling is desirable. We stress that in our model there are two pieces of potentially relevant information: trait and quality. While the former can only be disclosed by the seller, paying a disclosure cost, the latter can only be acquired by the buyer, paying an effort cost. We first conduct the analysis assuming that the joint distribution of quality and traits is exogenous, and then we extend the analysis endogenizing the distribution.

The remainder of the paper is organized as follows. In Subsection 1.2 we summarize our main results and provide their underlying intuition. In Subsection 1.3 we describe an application of the model to food products to help interpreting both model and results. In Subsection 1.4 we review the related literature, indicating in what respects our model is different from existing models. In Section 2 we first describe the persuasion game with labelling, assuming an exogenous joint distribution of qualities and traits (Subsection 2.1), and we then analyze the model and characterize the pooling equilibrium with no disclosure, showing that it survives D1 (Subsection 2.2). In Section 3 we study the effects of mandatory labelling, firstly by assuming an exogenous distribution of qualities and traits (Subsection 3.1), and then considering a variant of the model where the joint distribution of qualities and traits is endogenous (Subsection 3.2). Section 4 concludes, discussing some ideas for future research. Proofs are collected in the Appendix.

1.2 Our contribution

Our main contribution is the identification of a novel reason why mandatory labelling can backfire, which is different from the one pointed out by Jovanovic (1982), where inefficiencies

[4] To avoid possible misunderstandings, we clarify that our space of traits is one-dimensional, with the two traits being actually two levels along the same dimension.

[5] An alternative term for "labelling" is "certification", which in some applications is more common (see Subsection 1.3 about food products and GMO certification).

arise because of positive disclosure costs that outweigh the benefits of disclosure. Indeed, the possibility that mandatory labelling backfires is due to the fact that it reduces the incentive of buyers to incur a cost to learn quality, and it does not vanish when disclosure costs tend to zero. Furthermore, we find that the negative effects of mandatory labelling are borne by high quality producers, whose profits decrease, while profits for low quality producers may well increase; remarkably, this can trigger a reduction of the average quality in the long-run, when firms are free to exit and enter the market. This result can be seen as a way to formalize the perverse effects on the producer side that mandatory disclosure can yield, as foreseen, e.g., by Loewenstein et al. (2014).

More precisely, our first result is that, under exogenous distribution of qualities and traits, market unravelling can fail. Voluntary labelling only emerges if labels are sufficiently informative on quality and the cost to acquire information for buyers is quite large. On the contrary, whenever the cost to acquire information is small, a pooling equilibrium exists where no seller's type discloses his own trait. Importantly, this equilibrium turns out to be robust to a strong refinement as the D1 criterion (Cho and Kreps, 1987). The reason is that, thanks to the possibility of acquiring information on actual quality by some buyers, low quality sellers have a systematic incentive to disclose their trait in the hope of preventing information acquisition; this in turn induces buyers to believe that a seller deviating from a non-disclosure equilibrium is a seller of low quality.

We take this pooling equilibrium as the starting point to assess the desirability of mandatory labelling. In particular, we investigate the consequences of mandatory labelling by considering the effects of compulsory disclosure of sellers' traits. We find that mandatory labelling can potentially benefit buyers – if it allows them to save on the acquisition cost – but it can backfire on the seller's side – profits can increase for low quality sellers, and always decrease for high quality sellers.[6] This happens whenever the observation of the label – which helps buyers to have more precise beliefs about product quality – crowds out the buyer's incentive to costly acquire quality. This is an important observation as it suggests that, when the distribution of qualities and traits is determined by some form of competition among different seller's types, mandatory disclosure might have perverse consequences on average quality, sellers' profits and buyers' utility.[7]

[6] This contrasts with the finding in Bertomeu and Cianciaruso (2018), where they provide a quite general setting in which any type of sender cannot benefit from an external commitment to reduce discretion; we guess that such contrast comes from having no convexity in beliefs in our model (as defined in Bertomeu and Cianciaruso, 2018), due to the non-monotonicity in beliefs of buyer's optimal decision whether to acquire information.

[7] Schmeiser (2014) studies mandatory disclosure on a different piece of information: the relative impor-

To understand what happens when the competitive pressure induces adjustments in the distribution of qualities and traits, we modify the model by letting the distribution of seller's types be endogenous and, in particular, determined by a condition of equi-profitability across the seller's types that stay on the market. We find that, in this new setup, there can be a multiplicity of equilibria. Given the fact that we are studying a situation where a public authority can impose mandatory labelling, it seems reasonable to assume that the same authority can also subsidize sellers for a while in order to move the market to the desired equilibrium, which then should be self-enforcing. So, as an equilibrium selection criterion, we opt to focus on the equilibrium where average quality is the highest, because it maximizes both buyers' utility and seller's profits (which might be understood as the objective of the public authority) and, in addition, it is stable under reasonable profit-monotone dynamics. We show that, whenever the cost to acquire information is small enough, a pooling equilibrium exists where no seller's type discloses his own trait, i.e., market unravelling fails. Moreover, we find that, when we start from such equilibrium, imposing mandatory labelling can not lead to an increase in average quality nor to an increase in buyers' utility. Further, if average quality is successfully kept at the pre-labelling level, then seller's profits can not be larger and cost inefficiencies necessarily arise, net of disclosure costs. In short, mandatory labelling is confirmed to be potentially detrimental.

The main intuition for our results points to a crowding-out effect on the buyers' effort to learn quality. Labelling provides buyers with some extra information on quality. This is good for buyers, but it also creates an incentive not to acquire more precise information on quality, since the acquisition is costly. If buyers indeed stop acquiring information on quality, then low quality sellers can make greater profits, especially those having the trait that correlates more with quality. So, very informative labels can induce a more severe problem of asymmetric information, damaging the profits of high quality sellers who, if the distribution of seller's types is endogenous, progressively reduce their presence on the market. When the average quality of products with the same label is sufficiently low – at least as low as the overall average quality before the imposition of mandatory labelling – the label is not anymore very informative, so that buyers acquire again information on quality, boosting the profits of high quality sellers and so preventing a further decline in average quality. This, however, must occur for each given label in equilibrium. Hence, the following kind of cost-inefficiency can be brought about by mandatory labelling: the compulsory distinction of goods based on their labels prevents the possibility that high quality sellers specialize in

tance of product attributes. In such setup, he shows that inferential mistakes can lead to over or under-regulation by regulators and over or under-estimation of the importance of product attributes by consumers.

one trait and low quality sellers specialize in the other trait, even when this is cost efficient.

1.3 An application to food products

The theoretical analysis presented in this paper is quite abstract as it is meant to apply to different situations across different markets. For the sake of concreteness, and in order to provide an example that helps to interpret the model, we sketch here a simple application to the food market that we find relevant and of general interest. We use it to illustrate both the fitting of the model to a concrete case and the resulting insights in terms of the desirability of disclosure.

Consider markets for food, ranging from raw vegetables to complex industrial food that is ready to be eaten. Think of, for instance, the certification of the trait "organic", as opposed to "non-organic". The organic label is gaining increasing attention by consumers as it is believed to convey information on food quality. Obtaining the organic label is quite costly for a producer, but nevertheless it is nowadays widespread and currently expanding. This is a case of voluntary trait disclosure. Note that consumers are typically not interested in the organic trait per se, but rather they are interested in taste or healthiness of products. Also, it seems fair to say that the food products labelled as organic are not at all homogeneous in terms of average taste or healthiness, ranging from very healthy raw vegetables to not very healthy industrial sweets with some organic ingredients (enough to get the label), and from high quality organic wines to fancy bottled wines that have obtained the organic label, but whose quality is quite poor (apart from being low in sulfites).

So, why firms do not directly disclose the healthiness of a food product or its taste? One reason is that what is healthy for a consumer in part depends on her own needs and beliefs, making the actual healthiness of a product something that is individual-specific. Indeed, given an individual's current health status, it may be more important to have a low consumption of saturated fats relatively to unsaturated fats, or to guarantee a sufficient intake of specific vitamins or mineral salts, or to follow a calorie restriction diet. Moreover, different individuals may differ in what they think is a proper nutrition regime: think of the variety of commonly adopted dietary restrictions (e.g., mediterranean, zone, raw food, vegetarian, vegan, halal, kosher, etc.). A common reason why firms do not disclose the taste of a product is that it cannot be objectively certified. Besides being subjective, taste has several qualitative dimensions that are extremely hard to quantify in a verifiable way (e.g., bitterness, body or astringency for wine). In this case a buyer can develop an expertise that allows him to assess taste before he finalizes the purchase. Indeed, sometimes a taste expert is used by sellers to signal tastefulness (e.g., wine evaluations by sommeliers), but while such signals are often informative on taste for everybody, they are definitely imperfectly so because many other factors can affect a subjective assessment that produces no hard information. In either case, a buyer who carefully elaborates the available information on the product can potentially learn its actual healthiness or taste, that is, more than what the seller can certifiably disclose.

For many other potential traits of food products we do not observe voluntary certification, but either no certification at all or mandatory certification. One reason for the lack of certification for a given trait (e.g.,"easy-to-digest") is that the interested buyers are good enough at finding by themselves what is the product trait (e.g., what they digest better), and so producers need not invest in certifying their products. In terms of the model of this paper, this amounts to have buyers who are able to learn the trait at very low cost, which makes voluntary (and mandatory) certification useless. Another reason for the lack of voluntary certification is that enough buyers are quite good at finding out by themselves what is the quality of the food they consider to buy (e.g., say freshly fished fish at seaports). In terms of the model of this paper, this amounts to having buyers with very low cost of learning quality, which makes voluntary certification of a correlated trait, (e.g., the "freshness" measured by the day/hour of fishing and maybe conservation method) undesirable for high quality producers. Here mandatory labelling, if technically feasible, could backfire in that it might lead some buyers to just rely on the trait instead of actively learning quality, resulting in lower willingness to spend and possibly shifting profits from high quality producers to low quality producers.

The backfiring of mandatory labelling might have actually operated in the sense described above for the case of mandatory labelling of genetically modified organisms (GMO). Even not taking for granted that GMOs are unhealthy, it seems fair to say that the GMO-free label has been used by many consumers as an indication of food healthiness. Now, it is a fact that in the class of GMO-free food there is a great variance in terms of food healthiness, so that the fact that consumers may have relied on the label to make their purchases could have increased the profits of producers of unhealthy GMO-free food at the expenses of producers of healthy food that necessarily contains some GMO (e.g., soja-based products).[8] Another example of mandatory disclosure that in some cases might have backfired is the certification of food "origin". Some regulations impose for certain products to certify their area of origin

[8] As reported, e.g., by Albert (2010), there is little consensus on labelling products which do not contain any GM material but were derived from a GM crop or labelling because of the process of production; this uncertainty leaves room for profit-maximizing firms to save on materials and processes which, even if involving genetic modification to some extent, are not strictly regulated by the GMO legislation.

(from the region at a country level, up to the single piece of land from which the raw food used as input is obtained). Origin is often a trait which is informative about quality and consumers can rarely certify it by themselves. As a result, a certification of origin which is recognized as good could lead consumers to renounce acquiring further information on the product, increasing the profits of producers of low quality food with good origin and the expenses of producers of high quality food with bad origin.

1.4 Related literature

This paper contributes to three closely related streams of literature regarding the disclosure of product information to a potential buyer.

The first stream is on persuasion games, where a seller can provide verifiable information to a buyer in order to influence her actions (Milgrom and Roberts, 1986).[9] Persuasion games are different from cheap talk games where all reported information is unverifiable (Crawford and Sobel, 1982). In cheap talk games, when the buyer's optimal action is unique in the seller's type, persuasion is attained with full revelation of private information if and only if there is no seller's type that strictly prefers to be misidentified for another. Instead, in a standard persuasion game – where the seller can certify the disclosed information at no cost – persuasion requires that all information is revealed in equilibrium (Milgrom, 2008). This outcome crucially relies on the possibility for the buyer to have skeptical or pessimistic beliefs, in the sense that non-revelation has to be thought of as due to unfavorable private information (see Seidmann and Winter, 1997, for a generalization of this result that does not rely on seller's preference monotonicity).[10] Our model is a variant of a persuasion game where the seller can incur a cost to reveal and certify information – i.e., the seller can provide a certified label – which is correlated with product quality, whereas the buyer observes the behavior of the seller and then decides whether to exert effort in order to acquire information about quality on her own. In our setup, beliefs can turn out to be optimistic since low quality sellers often gain more from disclosing their private trait than high quality sellers.[11]

[9] The literature on Bayesian persuasion (Kamenica and Gentzkow, 2011; Kamenica, 2019) considers persuasion activities in a setting that differs from persuasion games since the sender can ex ante choose to be imperfectly informed.

[10] Giovannoni and Seidmann (2007) show that if the seller has the ability to certify all subsets of types containing the realized one, then there exists a fully revealing separating equilibrium if and only if no pair of types strictly prefer to be misidentified for another.

[11] Other variants of persuasion games have been studied. Anderson and Renault (2013), building on Anderson and Renault (2006), extend the persuasion game by allowing for search characteristics by consumers (as opposed to the experience characteristics treated in Milgrom, 1981). They generally confirm that the

The second stream of literature is more focused on market disclosure and on certification costs (see Dranove and Jin, 2010, for a survey covering also the empirical side). The most important finding in this literature is probably the so called "market unraveling": the best quality seller is the first to disclose in order to distinguish himself from lower quality sellers, generating an incentive to do the same for the second best seller, and so on and so forth. Importantly, Grossman (1981) and Milgrom (1981) show that, if there is no cost to disclose and certify information on quality, then sellers will always disclose in equilibrium. Again, this happens because buyers have skeptical beliefs: if no information is disclosed buyers infer that non-disclosing sellers are of low quality. As a consequence, sellers will voluntarily disclose their private information on quality with the result that mandatory disclosure is not necessary.[12] Instead, when disclosure is costly to the seller, Grossman and Hart (1980) and Jovanovic (1982) show that, in equilibrium, only sellers with product quality above a cost-dependent threshold disclose.[13,14] Matthews and Postlewaite (1985) and Shavell (1994) introduce a cost for the seller to acquire information on own quality and show that in such a case mandatory disclosure may motivate sellers to reduce information collection, hence backfiring.[15] Our model introduces a cost for the *buyer* to acquire information on quality, which allows for the existence of pooling equilibria where sellers do not disclose their private information. Such equilibrium is sustained by out-of-equilibrium beliefs that punish the deviating sellers: an unexpected observation of a label reasonably leads buyers to believe that they are in front of a low quality seller, because in the absence of a label buyers acquire information on their own and this makes a high quality seller less likely to gain from disclosing his trait. As the introduction of mandatory disclosure makes such reasoning impossible – since all sellers are obliged to disclose their traits – it can result in an advantage for low

outcome is a separating equilibrium with quality unravelling, but they also show that unravelling may fail for low enough search costs.

[12] Koessler and Renault (2012) show that, under mild assumptions, unraveling obtains also when information disclosure is possible on horizontal attributes as well as on quality.

[13] In an empirical analysis, Butler and Read (2017) point out that full information revelation fails in the hospitality industry, finding instead a downward linear relationship between quality and disclosure. Also, based on evidence from the German housing market, Frondel et al. (2020) conclude that mandatory disclosure is more effective than voluntary disclosure.

[14] A similar result is obtained by Okuno-Fujiwara et al. (1990), who show that unraveling can be upset when information cannot be fully disclosed, resulting in a failure of perfect revelation.

[15] Recent contributions to the literature on market disclosure (Cheong and Kim, 2004; Board, 2009; Levin et al., 2009; Sun, 2011; Hotz and Xiao, 2013; Celik, 2014) show that competition among multiple sellers can prevent disclosure even if the disclosure cost is zero. Emons and Fluet (2012) show that when comparative disclosure is available a firm advertises the quality differential. Janssen and Roy (2014) show that when prices can convey information on quality, competition can make them a substitute for certified disclosure.

quality sellers.[16]

The third stream of literature regards market disclosure when buyers are not perfectly able to gather or understand information (see Loewenstein et al., 2014, for a recent survey covering also the psychological literature, and Mengel, 2012, on the evolutionary selection of coarse information partitions).[17] For instance, buyers can be unable to understand the information disclosed (Fishman and Hagerty, 2003) or be unaware of disclosures made by sellers (Dye and Sridhar, 1995). In such cases, market unraveling might fail even if disclosure costs are negligible (Gabaix and Laibson, 2006). Li et al. (2014) show that a larger share of unaware consumers makes information disclosure less likely to occur and mandatory disclosure more likely to be optimal. Further, psychological evidence suggests that people do not fully decide how to allocate attention, mostly focusing on salient product features and disregarding other features, even if they are relevant (Bordalo et al., 2013; Kalaycı and Serra-Garcia, 2015). Kiesel and Villas-Boas (2013) find experimentally that nutritional labels reduce consumer's cost to acquire information on product quality, affecting consumer behavior. Our model is related to this literature in that we also consider consumers who have cognitive bounds. However, instead of dealing with such bounds as a reason for suboptimal choices, we maintain our analysis within the standard economic framework where individuals make optimal choices. Indeed, we model cognitive bounds as a cost that the buyer has to incur in order to process available information on quality. In particular, we follow Dewatripont and Tirole (2005) and, more closely, Bilancini and Boncinelli (2018a), where a stylized model is provided that is inspired by dual-process theories of information elaboration.[18] Indeed, the buyers' decision whether to acquire or not information on quality can be seen as a choice between two different cognitive routes: one cheap and fast, which does not require much effort but does not allow to learn actual quality, and one costly and slow, which requires effort but allows to obtain the desired information.[19] This distinguishes our approach from models as in Fishman and Hagerty (2003) where buyers cannot choose

[16] Even if we have focused on sellers who are interested to signal their quality to consumers in order to obtain an appropriately high price for their product, demand for quality certification may also come from buyers, who do not want to overspend on low quality (see Stahl and Strausz, 2017, for a comparison between seller-induced certification and buyer-induced certification).

[17] Bilancini and Boncinelli (2018c) study costly acquisition of signals, pointing out how this makes pooling a more prominent outcome.

[18] See, e.g., Evans and Stanovich (2013) for an overview of the psychological approach to dual process reasoning.

[19] The cognitive limitations arising from the cheap and fast cognitive route are modeled by Bilancini and Boncinelli (2018b) as analogical reasoning within a signaling model.

to acquire more information at a cost.[20]

Recently, a few studies have started investigating the strategic disclosure of information in an experimental setting, providing evidence for both incomplete unraveling and receiver naivete. In a simple two-person disclosure game, Jin et al. (2020) find that senders disclose favorable information, but withhold less favorable information. Deversi et al. (2018) also find a strategic use of vagueness of information: when news are unfavorable they are often delivered under the disguise of vagueness; indeed, some (naive) receivers are systematically misled by it.

2 Model and preliminary results

2.1 A persuasion game with labelling

A seller, denoted by S, wants to sell its products to a buyer, denoted by B. (We will sometimes refer to S as "he" and to B as "she".) The quality of S's product is $q \in \{H, L\}$ where H denotes high quality and L denotes low quality; moreover, the product has a certifiable characteristic or trait $t \in \{X, Y\}$ which is known to be correlated with product quality. Initially, B ignores both q and t, but she knows that S's type is one of the four possible combinations of quality and trait, i.e., $(q, t) \in \{H, L\} \times \{X, Y\}$.

We denote with $p(H)$ the prior probability that $q = H$, i.e., $(q, t) = (H, t)$, $t \in \{X, Y\}$. Also, $p(L) = 1 - p(H)$ is the prior probability that $q = L$. We further (and crucially) assume that the trait t is informative about quality, namely that X is positively correlated with H while Y is positively correlated with L (and, hence, negatively correlated with H). Formally, if we denote with $p(q|t)$ the probability of quality being q conditional on trait being t, we impose that $p(H|X) > p(H|Y)$, which implies $p(L|X) < p(L|Y)$, $p(H|Y) < p(H) < p(H|X)$, and $p(L|X) < p(L) < p(L|Y)$. Finally, let $p(X)$ and $p(Y) = 1 - p(X)$ denote, respectively, the prior probability that $t = X$ and $t = Y$.

The trait t can not be directly observed by B, but S can incur the cost $c_d > 0$ to disclose t and certify it. In particular, when a type (q, t) pays c_d, B learns t, and can update her beliefs on q accordingly. The buyer can exert effort and incur the cost $c_e > 0$ to learn q. After possibly observing trait and/or quality, each buyer chooses some scalar z, which can

[20] Perhaps closer to our model are those studies where consumers can acquire quality information by incurring some cost, such as Bar-Isaac et al. (2010, 2012), in which firms invest in quality to induce the desired information acquisition by consumers, and Wang (2013), where where firms use advertisement to deter consumers' search.

be interpreted as the amount of money that she spends on the firm's products.[21]

More precisely, we follow Milgrom (2008) in the interpretation of the scalar z. On the one hand, we can think of products whose prices are fixed due to regulation by a public authority; here z is the quantity that the buyer chooses to purchase at the given price. On the other hand, we can think of z as the highest price that the buyer is willing to pay for a unit of product; in this interpretation, a higher z should allow the seller to obtain higher profits, for instance thanks to some form of price discrimination as discounting. Carrying on the example on food products used in Subsection 1.3, we stress that price regulation is not so rare in food markets, and periodic discounting is quite a common practice for food companies.

We assume that the buyer's payoff (gross of acquisition costs) is $U(z,q)$, with $\partial U(0,q)/\partial z > 0$, $\partial U(k,q)/\partial z < 0$ for some $k > 0$, and $\partial^2 U(z,q)/\partial z^2 < 0$ (an optimal choice exists and is positive, and marginal utility of z is decreasing), and $\partial U(z,H)/\partial z > \partial U(z,L)/\partial z$, which means that the marginal value of an increase in z to the buyer is increasing in quality q.[22] The seller's payoff (gross of disclosure costs) is denoted with $V(z)$, with $\mathrm{d}V(z)/\mathrm{d}z > 0$, which means that the seller always prefers the buyer to choose a higher z.[23] Both the seller and the buyer maximize the expected payoff. We denote with $\mu(H)$, $\mu(H|x)$ and $\mu(H|y)$ the generic beliefs maintained by B when she observes, respectively, no trait, trait x, trait y.

Summing up, a strategy for S is represented by function a $\sigma : \{H,L\} \times \{X,Y\} \to \{0,1\}$ mapping a type (q,t) into either the choice of disclosing his own trait t, which is denoted by 1, or not disclosing it, which is denoted by 0. A strategy for B is represented by a function $\beta : \{0,x,y\} \to \{\underline{e}\} \times \mathbb{R}_+ \cup \{\overline{e}\} \times \mathbb{R}_+^2$ mapping each possible information about the observed trait – 0 for no trait, x for trait X, y for trait Y – into the choice of whether to exert the

[21] Our main results on the effects of mandatory labelling are not crucially dependent on z being a scalar, which is only used to provide a simpler analysis. In case z can only assume finitely many values, some of the arguments used in the analysis should be refined to consider that the buyer's optimal z would not be anymore strictly increasing in the buyer's beliefs that quality is high, but only weakly increasing: indeed, a larger belief does strictly benefit the seller only if such increase translates into a higher z.

[22] As remarked by Milgrom (2008), this latter assumption is not entirely general: consumers could spend less on higher quality products when, e.g., quality means a reduced need for replacement. However, we note that if $\partial^2 U(0,q)/\partial z \partial q < 0$ then results would still apply but with types inverted, as firms gain more by being recognized as L types.

[23] In some cases it is reasonable to think that seller's profits depend on q as well, as producing high quality can be more expensive than producing low quality. In addition, profits might depend on the trait either. All these are possible generalizations of the model that intuitively do not change the quality of our main results. Non-trivial complications might arise if a welfare analysis is conducted, since differences in surplus across pairs of quality and trait should be taken into account.

effort to acquire q, denoted by $\overline{e}$, and how much to spend depending on the quality observed (L or H), denoted by $(z_L, z_H) \in \mathbb{R}^2_+$, or not to exert effort, denoted by $\underline{e}$, and how much to spend independently of quality, denoted by $z \in \mathbb{R}_+$.[24]

Figure 1 provides a graphical summary of the timing of the game.

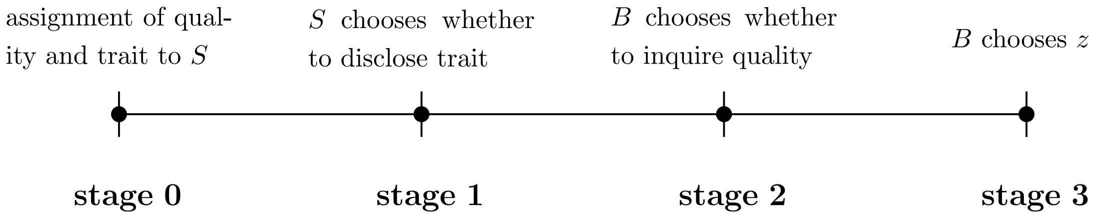
assignment of quality and trait to S
S chooses whether to disclose trait
B chooses whether to inquire quality
B chooses z
stage 0
stage 1
stage 2
stage 3

Figure 1: **Timing of the game**. At stage 0, the seller receives one of four possible combinations of quality and trait. At stage 1, the seller chooses whether to publicly disclose the trait, conditional on the quality and trait received. At stage 2, the buyer chooses whether to acquire knowledge of the quality, conditional on the observation of the trait, if disclosed. At stage 3, the buyer chooses z.

As a solution concept we focus on the *weak Perfect Bayesian Equilibrium* (wPBE) in pure strategies.[25] Given the large variety of equilibria typically arising in signaling games, various kinds of restrictions on out-of-equilibrium beliefs have been used as refinements. In this paper, we will rely on the D1 criterion (Cho and Kreps, 1987) with the purpose of limiting the analysis to equilibria that can be considered rather robust. Basically, D1 imposes that, if B observes a deviation by S, then B puts zero probability on any type of S whose set of beliefs justifying such deviation is strictly contained in the set of beliefs justifying such deviation for another type.

2.2 Pooling equilibrium with no voluntary labelling

We start by analyzing the optimal behavior of the buyer. Suppose that after observing either trait x, or trait y, or 0 (i.e., no trait has been disclosed), the buyer has formed a belief μ

[24] This distinction can be traced to the classification of elaboration processes known in psychology as *dual process reasoning*: while "System 1" is fast, cheap and intuitive, "System 2" is slow, costly and analytical (see, e.g., Kahneman, 2003). We stress this interpretation based on cognitive effort because we think that it well applies to many cases of information acquisition of products quality. Of course, other interpretations are possible where the cost of acquiring information on quality is entirely due to non-psychological factors.

[25] A wPBE basically requires sequential rationality and consistency of equilibrium beliefs, while out-of-equilibrium beliefs are left totally free (see Mas-Colell et al., 1995, Definition 9.C.3).

that the product is of high quality. If B chooses $\underline{e}$ (i.e., she does not exert effort to acquire information on quality), then her expected payoff as a function of z is $\mu U(z,H) + (1-\mu)U(z,L)$, whose maximum is reached at z^*_μ with $\mu \partial U(z^*_\mu, H)/\partial z + (1-\mu)\partial U(z^*_\mu, L)/\partial z = 0$. If B chooses $\overline{e}$ (i.e., she exerts effort to acquire information on quality), then her expected payoff as a function of z is $\mu U(z_H, H) + (1-\mu)U(z_L, L) - c_e$, whose maximum is reached at z^*_H and z^*_L with $\partial U(z^*_H, H)/\partial z = 0$ and $\partial U(z^*_L, L)/\partial z = 0$.

In order to understand what is the optimal buyer's behavior, we have to compare $\mu U(z^*_\mu, H) + (1-\mu)U(z^*_\mu, L)$, which is the maximum expected payoff earned if $\underline{e}$ is chosen, and $\mu U(z^*_H, H) + (1-\mu)U(z^*_L, L)$, which is the maximum expected payoff earned if $\overline{e}$ is chosen. This leads us to the following lemmas:

Lemma **1.** *(Optimality of $\underline{e}$ for B)*
For any given acquisition cost c_e, there exist $\underline{\mu}(c_e)$ and $\overline{\mu}(c_e)$ such that, if $\mu \in [0, \underline{\mu}(c_e)] \cup [\overline{\mu}(c_e), 1]$, then $(\underline{e}, z^*_\mu)$ is an optimal response for the buyer. In addition, if $\mu \in [0, \underline{\mu}(c_e)) \cup (\overline{\mu}(c_e), 1]$ then it is the only optimal response.

Lemma **2.** *(Optimality of $\overline{e}$ for B)*
For any given belief on quality μ, there exists $\hat{c}_e(\mu)$ such that, if $c_e \leq \hat{c}_e(\mu)$, then $(\overline{e}, z^*_L, z^*_H)$ is an optimal response for the buyer. In addition, if $c_e < \hat{c}_e(\mu)$ then it is the only optimal response.

Lemmas 1 and 2, together with a concavity argument on the difference between the expected utility of choosing $\overline{e}$ and $\underline{e}$, allow us to obtain the following proposition on the buyer's optimal choice as a function of μ:

Proposition **1.** *(Optimal choice for B)*
There exists $\hat{c}_e$ such that, if $c_e < \hat{c}_e$, then there exist $\underline{\mu}(c_e)$ and $\overline{\mu}(c_e)$, with $\underline{\mu}(c_e) < \overline{\mu}(c_e)$, such that, if $\mu \in [0, \underline{\mu}(c_e)) \cup (\overline{\mu}(c_e), 1]$, then $(\underline{e}, z^*_\mu)$ is the only optimal response for the buyer, while if $\mu \in (\underline{\mu}(c_e), \overline{\mu}(c_e), 1)$, then $(\overline{e}, z^*_L, z^*_H)$ is the only optimal response for the buyer.

The results in Proposition 1 can be summarized graphically as in Figure 2. From the figure we can understand that, when the beliefs on quality are low or high – respectively, $\mu \in [0, \underline{\mu}(c_e)]$ and $\mu \in [\overline{\mu}(c_e), 1]$ – then it is optimal for the buyer not to acquire knowledge of the actual quality of the product, saving on the acquisition cost. In these ranges of beliefs the optimal amount z^*_μ is increasing in the belief μ. Indeed, when the belief is rather extreme – i.e., sufficiently close to either 0 or 1 – the expected gain for the buyer to choose $\overline{e}$ over $\underline{e}$ is quite small (since the uncertainty is low), and hence the buyer finds it optimal not to pay the acquisition cost. However, when the belief is intermediate – i.e., $\mu \in [\underline{\mu}(c_e), \overline{\mu}(c_e)]$ – the

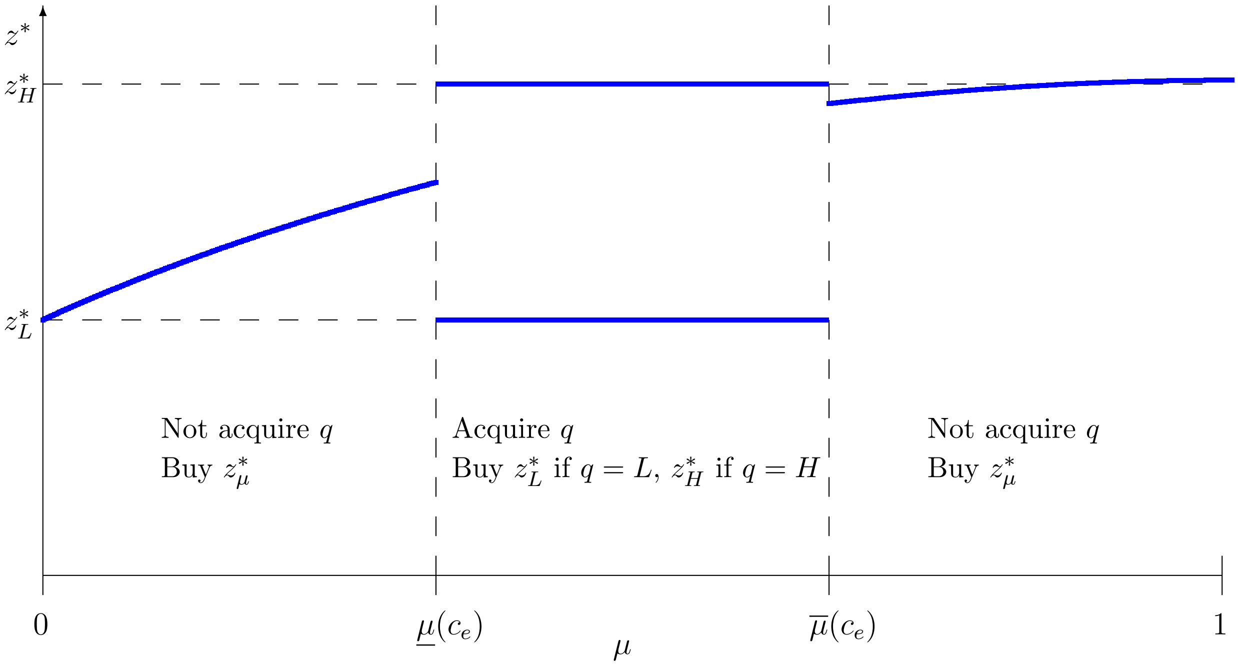

Figure 2: Buyer's optimal behavior as a function of beliefs μ on quality. When beliefs are rather extreme the buyer does not acquire information on quality. When beliefs are in an intermediate range, the buyer acquires information on quality, and the amount spent depends on the quality observed.

buyer finds it optimal to pay the acquisition cost c_e and acquire more precise knowledge of quality. In this range of beliefs the optimal amount z^* will be conditional on q, and it will be equal to either z_L^* or z_H^*. The existence of such an intermediate range of beliefs is ensured when c_e is low enough.

We now analyze the optimal behavior of the seller, whose expected gain from disclosing his private trait is very much dependant on the three beliefs that the buyer holds when she sees, respectively, the trait x, the trait y, and no trait. In general, a seller would like to disclose his private trait when the cost of disclosure c_d is more than compensated by an expected increase in z generated by inducing the buyer to have a higher belief on quality. In this regard, we note that the disclosure choice that turns out to be optimal for the seller can in principle differ for a high quality seller and a low quality seller, because in case the buyer chooses to acquire q she also chooses z conditionally on $q = H$ and $q = L$ (see Figure 2). In particular, if the buyer acquires q when no trait is observed, then disclosing t makes no sense for (H, X) and (H, Y), because they are recognized as high quality sellers with no need to pay the disclosure cost. Instead, disclosing t can in principle be profitable for (L, X)

and (L, Y), depending on the buyer's beliefs when t is observed.

We note that a variety of equilibria can arise in this setting, some of which are separating equilibria (where the disclosure of the private trait is a signal for quality) while others are pooling equilibria (where no signaling role of labelling emerges). Instead of discussing all these possible occurrences, we prefer to focus on a specific pooling equilibrium that is of interest for the subsequent analysis of mandatory labelling, that is, we focus on the pooling equilibrium where no trait is disclosed.

Proposition **2.** *(Pooling equilibrium with no voluntary disclosure)*
For any given prior belief $p(H)$, if $c_e < \hat{c}_e(p(H))$, then there exists a wPBE equilibrium satisfying the D1 criterion where no seller type discloses and the buyer acquires q choosing z^*_H if $q = H$ and z^*_L if $q = L$.

The intuition underlying Proposition 2 is straightforward. For $c_e < \hat{c}_e(p(H))$, the buyer prefers to acquire q and spend conditionally on it. Given this, type (H, X) maximizes his profits by not disclosing q, independently of the buyer's belief on quality. A fortiori, this holds for (H, Y). The fact that also L-types do not gain by disclosing depends on the buyer's out-of-equilibrium beliefs: in principle, if out-of-equilibrium beliefs triggered by deviating and disclosing t were high enough then $L-types$ might be more than compensated of the disclosure cost, given that L-types are recognized as such by the buyer in the profile under consideration. However, we may consider pessimistic beliefs, i.e., beliefs such that the disclosure of any t is interpreted as a signal of low quality. Such beliefs are particularly appealing in this setting, because L-type sellers always have a higher incentive to disclose t in the attempt to discourage the buyer from acquiring the information on quality (in the Appendix we show that pessimistic beliefs satisfy the D1 criterion). If out-of-equilibrium beliefs are pessimistic, disclosing t yields no benefit to L-types as well but still brings a cost, so that L-types do not gain by disclosing.

Let us conclude the section with one remark on what happens when the condition $c_e < \hat{c}_e(p(H))$ is not satisfied. For $c_e \geq \hat{c}_e(p(H))$, there are possible equilibria where X-types voluntarily disclose their trait. This happens if the buyer does not acquire q so that type (H, X) can find it profitable to disclose x, which in turn makes it profitable for type (L, X) too. We stress that the existence of such equilibria is not particularly relevant for our analysis. Since we want to assess the desirability of mandatory labelling, only equilibria where there is not full disclosure have to be considered – and Proposition 2 guarantees that they exist for reasonable parameter values. To put it differently, our analysis begins with the following observation: if the buyer is sufficiently reflective (in the sense that their c_e is

Player (type)	**Effect if:** $p(H\|X) > \overline{\mu}(c_e)$	**Effect if:** $p(H\|Y) < \underline{\mu}(c_e)$
seller (H, X)	$-$	/
seller (H, Y)	/	$-$
seller (L, X)	$+$	/
seller (L, Y)	/	$+$
buyer	$+$	$+$

Table 1: Effects of mandatory labelling on seller's types and buyer. When mandatory disclosure of trait X induces the buyer not to acquire the information on quality (i.e., if $p(H|X) > \overline{\mu}(c_e)$ for trait X, and $p(H|Y) < \underline{\mu}(c_e)$ for trait Y), high quality sellers are better off (i.e., sellers (H, X) and (H, Y)) and low quality sellers are worse off (i.e., sellers (L, X) and (L, Y)); also, the buyer is better off, saving on the cost of information acquisition.

sufficiently low) then market unravelling fails, so that mandatory labelling might be invoked (instead, if market unravelling obtains, then there is no need to evaluate the desirability of mandatory labelling).

3 The effects of mandatory labeling

3.1 Exogenous joint distribution of qualities and traits

Let us denote with (σ^p, β^p) an equilibrium profile that leads to a pooling outcome as described in Proposition 2. Also, let us denote with (σ^m, β^m) the equilibrium profile that results from the introduction of mandatory labelling, i.e., by imposing $\sigma^m(q, X) = x$ and $\sigma^m(q, Y) = y$ for all q.[26] We observe that under mandatory labelling the seller has no real choice to make, since σ^m is the unique feasible strategy; this eliminates the possibility of observing actions out of equilibrium, so that the buyer chooses β^m as follows: $\beta^m(x) = (\overline{e}, z^*_H, z^*_L)$ if $p(H|X) < \overline{\mu}(c_e)$, $\beta^m(x) = (\underline{e}, z^*_{p(H|X)})$ if $p(H|X) > \overline{\mu}(c_e)$ (being indifferent between $\underline{e}$ and $\overline{e}$ if $p(H|X) = \overline{\mu}(c_e)$), and $\beta^m(y) = (\overline{e}, z^*_H, z^*_L)$ if $p(H|X) > \underline{\mu}(c_e)$, $\beta^m(y) = (\underline{e}, z^*_{p(H|Y)})$ if $p(H|Y) < \underline{\mu}(c_e)$ (being indifferent between $\underline{e}$ and $\overline{e}$ if $p(H|Y) = \underline{\mu}(c_e)$).

By comparing (σ^p, β^p) with (σ^m, β^m), we can analyze the consequences of mandatory

[26] One might add a cost of certification or auditing that must be incurred by the authority. This extra cost does not change the quality of results, it simply makes mandatory disclosure relatively less desirable.

labelling. We first consider H types. One straightforward effect is that they must incur the cost c_d.[27] Further, since in (σ^p, β^p) the high types are selling z^*_H to the buyer, there is another potential negative effect: the buyer can be induced to use $\underline{e}$, which reduces her spending from z^*_H to $z^*_{p(H|t)}$, $t = X, Y$. For low types the net of positive and negative effects is ambiguous. Somewhat surprisingly, it turns out that low types can actually gain by mandatory labelling. A negative effect is given by the cost c_d, for both type (L, X) and type (L, Y). However, there is a potential positive effect for both seller's types. If the buyer's prior conditionally on observing x is sufficiently high as to induce the buyer to use $\underline{e}$, then type (L, X) increases sales from z^*_L to $z^*_{p(H|X)}$. A similar effect can arise for type (L, Y) if the buyer's prior conditionally on y is sufficiently low, but the increase in z^* is obviously lower for type (L, Y) than for type (L, X). So, type (L, X) is more likely to gain from mandatory disclosure than type (L, Y). The following proposition summarizes:

Proposition **3.** *(Effects of mandatory labelling on S)*
Let $c_e < \hat{c}_e(p(H))$. Starting from a pooling equilibrium with no voluntary disclosure, the imposition of mandatory disclosure of traits:

(i) lowers the payoff earned by the seller of types (H, X) and (H, Y);

(ii) raises the payoff earned by the seller of type (L, X) if $p(H|X) > \overline{\mu}(c_e)$ and $c_d < V(z^*_{p(H|X)}) - V(z^*_L)$;

(iii) raises the payoff earned by the seller of type (L, Y) if $p(H|y) < \underline{\mu}(c_e)$ and $c_d < V(z^*_{p(H|Y)}) - V(z^*_L)$.

Finally, mandatory disclosure affects the buyer's payoff positively, but the magnitude of this effect can vary substantially depending on how extreme are the buyer's beliefs, conditional on observing a label. There is a potentially positive effect due to the additional information conveyed by t: if $p(H|t)$ is very close to $p(H)$, then the observation of t leaves the buyer still too uncertain about quality, so that she prefers to keep using $\overline{e}$ (paying c_e and acquiring q); when $p(H|t)$ is sufficiently distant from $p(H)$ the buyer chooses to save c_e at the cost of a less precise assessment of quality. This latter effect is increasing in $p(H|X)$ and decreasing in $p(H|Y)$, i.e., it is increasing in the amount of information conveyed by the labels. The following proposition summarizes.

Proposition **4.** *(Effects of mandatory labelling on B)*
Let $c_e < \hat{c}_e(p(H))$. Starting from a pooling equilibrium with no voluntary disclosure, the imposition of mandatory disclosure of traits:

[27] The cost of disclosure can be totally or partly subsidized by the authority imposing the disclosure. This moves the burden from sellers to the taxpayers that finance the subsidy.

(i) leaves the buyer's payoff unchanged if $p(H|X) < \overline{\mu}(c_e)$ *and* $p(H|Y) > \underline{\mu}(c_e)$*:*

(ii) raises the buyer's payoff if $p(H|X) > \overline{\mu}(c_e)$ *or* $p(H|Y) < \underline{\mu}(c_e)$*, and the buyer's gains increase monotonically in* $p(H|X)$ *and decrease monotonically in* $p(Y|H)$*.*

Table 1 provides a qualitative summary of all the effects generated by mandatory disclosure for both the seller and the buyer.

3.2 Endogenous joint distribution of qualities and traits

The main message that can be extrapolated from the results presented in the previous subsection is the following: if we start from a situation where all seller's types do not voluntarily disclose their traits (as in the weak Perfect Bayes Nash equilibrium described in Proposition 2), then the introduction of mandatory disclosure of the private trait has the consequence of lowering profits for high quality firms and, if labels are sufficiently informative on quality, of increasing profits for low quality firms (Proposition 3). The buyer's payoff, instead, is not reduced by mandatory disclosure, and is actually increasing when labels are informative enough (Proposition 4). These effects, however, are obtained for given $p(H)$, $p(H|X)$ and $p(H|Y)$, i.e., when the distribution of seller's types is exogenous. When instead the distribution of seller's types is endogenous, further and possibly different effects may arise. In this section we investigate such effects.

In order to endogenize the distribution of seller's types, we adjust the model of Section 2 introducing heterogeneous setup costs for the different seller's types and letting $p(H)$ and $p(H|t)$ to be determined by a condition of equal profitability across the seller's types that stay on the market. Here we rely on the idea that in the absence of entry and exit barriers, we should observe over time an increase in the proportion of high quality firms if their profits are larger than those of low quality firms, or a decrease of such proportion in the opposite case. Therefore, equal profitability across the seller's types is expected to hold in the long run, after the above sketched adjustment process has come to a rest. In particular, we allow for the possibility that one or more types do not operate at all, and we denote with $p(H|t) = \emptyset$ the case where neither type (H, t) nor type (L, t) is on the market. We also allow for mixed behavior of the buyer, in order to guarantee equilibrium existence.[28]

A seller has to decide whether to produce or not, since setup and maintenance costs are not sunk. Moreover, a seller can choose which pair of quality-trait to produce, which entails different costs. Denote with c_{qt} the fixed cost that must be incurred for producing as type

[28] The possibility of mixed behavior of the buyer does not play any significant role in the model of Section 2, so one can safely neglect it.

(q,t). Let high quality be more costly to produce than low quality, i.e., $c_{HX} > c_{LX}$ and $c_{HY} > c_{LY}$. Also, in order to justify potential correlation between trait and quality, let trait X be relatively more complementary to the production of quality H and trait Y be relatively more complementary to the production of quality L, i.e., $c_{HX} < c_{HY}$ and $c_{LX} > c_{LY}$. Finally, we assume that $V(z^*_H) - c_{LY} > V(z^*_L) - c_{HY}$, i.e., high quality is always worth producing if q is public information.

The buyer's optimal behavior is always given by Lemmas 1 and 2, and Proposition 1. So, for any belief μ, the buyer's optimal behavior can be parsimoniously described by z^*_μ and the probability of acquiring q, which we denote with λ^* and must be equal to 0 when $\mu < \underline{\mu}$ or $\mu > \overline{\mu}$, equal to 1 when $\mu \in [\underline{\mu}, \overline{\mu}]$, and belonging to $[0,1]$ when $\mu = \underline{\mu}$ or $\mu = \overline{\mu}$.

When the buyer's beliefs on quality are obtained by means of Bayes' rule and the buyer makes optimal decisions, a seller of type (q,t) obtains the following profits:

$$\pi_{qt} = \begin{cases} \lambda^* V\left(z^*_q\right) + (1-\lambda^*) V\left(z^*_{p(H)}\right) - c_{qt} & \text{if } t \text{ is not disclosed} \\ \lambda^* V\left(z^*_q\right) + (1-\lambda^*) V\left(z^*_{p(H|t)}\right) - c_{qt} - c_d & \text{if } t \text{ is disclosed} \end{cases} \tag{1}$$

In this setup, we adjust the definition of a profile to allow for mixed behavior in terms of effort exertion. In particular, we denote a profile with $(\sigma, \tilde{\beta})$, where σ describes the choice of disclosure for each seller's type (as defined in Section 2), and $\tilde{\beta} : \{0, x, y\} \to [0,1] \times \mathbb{R}_+ \times \mathbb{R}^2_+$ maps each possible information about the observed trait into a probability $\lambda \in [0,1]$ to exert effort and acquire q, and into expenditure choices which can be conditional on quality only if effort has been exerted.

The only differences in this setting with respect to the persuasion game with labelling introduced in Section 2 are that (i) the buyer is allowed to choose probabilistically between $\underline{e}$ and $\overline{e}$, and (ii) some of the seller's types may not be present.

A *persuasion equilibrium with endogenous seller's types* (PEEST) is defined as a triple $(p(H),\ (p(H|X), p(H|Y)),\ (\sigma, \tilde{\beta}))$, such that: (i) $(\sigma, \tilde{\beta})$ is a wPBE[29] which survives the D1 criterion of the persuasion game induced by $p(H)$ and $(p(H|X), p(H|Y))$, and (ii) profits π_{qt} are equal for all types (q,t) such that $p(H|t) > 0$ if $q = H$ and $1 - p(H|t) > 0$ if $q = L$ and not greater for other types, i.e., all seller's types on the market must earn equal profits while those out of the market must not be capable of earning more. As done in Section 2, instead of discussing all possible kinds of equilibria that can arise, we focus on the kind of pooling PEEST that is of interest for the subsequent analysis of mandatory labelling. In

[29] The difference in the definitions of $\tilde{\beta}$ and β requires formal adjustments in the qualification as wPBE, which however entail no substantial difference in the analysis.

particular, we focus on the PEEST which induces the highest $p(H)$, and we refer to it as the *best PEEST*. The reason for this choice is that, typically, the public authority can not only impose mandatory labelling but also subsidize high quality producers for a while. So, the equilibrium with the highest $p(H)$ would always be selected, as it maximizes surplus. A potential problem could regard the stability of such equilibrium. However, under plausible profit-based dynamics, the equilibrium with the highest $p(H)$ is the only stable equilibrium such that $p(H) > 0$. Figure 3 gives an intuition of this.

It turns out that the best PEEST entails no disclosure, with a sort of "perfect correlation" between q and t emerging endogenously: only seller's types (H, X) and (L, Y) operate and average quality is neither too low nor too high as to induce the buyer to acquire q often enough, hence sustaining the profits of H-types. The following proposition summarizes:

Proposition 5. *(Best PEEST with no voluntary disclosure)*
There exists $\check{c}_e > 0$ such that, if $c_e < \check{c}_e$, then the best PEEST is $(\overline{\mu}(c_e), (1, 0), (\sigma^p, \tilde{\beta}^p))$, where no operating seller's type discloses his own trait, the acquisition probability is $\lambda^* = \frac{(c_{HX} - c_{LY})}{V(z_H^*) - V(z_L^*)}$, and expenditure choices are: z_H^* if $q = H$ is observed, z_L^* if $q = L$ is observed, and $z^*_{\overline{\mu}(c_e)}$ if quality remains unobserved.

The intuition underlying Proposition 5 is simple. No equilibrium with $p(H) > \overline{\mu}(c_e)$ can exist, because for such $p(H)$ no buyer would acquire q and this would wipe out of the market H-types. So, the PEEST with highest $p(H)$ can be at most such that $p(H) = \overline{\mu}(c_e)$. For this value of $p(H)$, the buyer is indifferent between acquiring q and not acquiring it; therefore, it can be found a probability of effort exertion, λ^*, that represents an optimal behavior for the buyer and allows H-types to remain on the market.[30] Further, λ^* can be large enough to discourage the disclosure of x by type (H, X), which in turn induces all seller's types not to disclose (see the discussion below Proposition 2). Then, under no disclosure, type (H, X) obtains the same gross profits of type (H, Y), but incurs lower setup costs since $c_{HX} < c_{HY}$; this leads type (H, Y) out of the market. Similarly, type (L, Y) obtains the same gross profits of type (L, X) but incurs lower setup costs since $c_{LY} < c_{LH}$, and this leads type (L, X) out of the market.

As done in Section 3 we model the introduction of mandatory labelling by imposing $\sigma = \sigma^m$, where $\sigma^m(q, X) = x$ and $\sigma^m(q, Y) = y$, for all q. Then, we study the consequences of mandatory labelling by looking at the effects on average quality, buyer's payoff and production costs. It turns out that mandatory labelling may often be undesirable. The following proposition summarizes:

[30] We note that it always exists a PEEST for $p(H) = \underline{\mu}(c_e)$. However, this equilibrium is intuitively not stable under reasonable profit-based dynamics, as depicted in Figure 3.

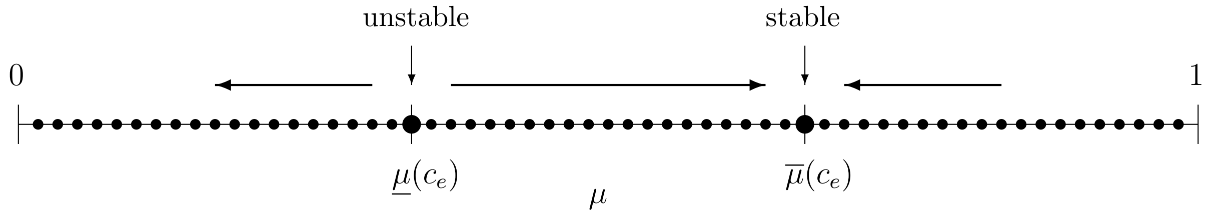

Figure 3: The dynamics of $p(H)$. Consider the case of no disclosure: for $p(H) > \overline{\mu}(c_e)$ or $p(H) < \underline{\mu}(c_e)$ no buyer acquires q, boosting the profits of L-types which, under reasonable profit-based dynamics, leads to a reduction in $p(H)$; for $\underline{\mu}(c_e) < p(H) < \overline{\mu}(c_e)$, the buyer always acquires q boosting the profits of H-types, and this leads to an increase in $p(H)$. A similar argument applies to $p(H|t)$ when x or y are observed.

PROPOSITION **6.** *(Effects of mandatory labelling)*
Starting from the best PEEST with no disclosure, when the distribution of seller's types is endogenous the imposition of mandatory disclosure of traits does not increase the average quality nor the buyer's payoff; also, if high quality is present in the market, it generates cost inefficiencies net of disclosure costs.

The results described in Proposition 6 can be explained as follows. As discussed for the intuition of Proposition 5, the maximum $p(H)$ that can be obtained with an endogenous distribution of types is equal to $\overline{\mu}(c_e)$, and this is true for any piece of information obtained (being it x, y or 0). In other words, the maximum average quality that is obtainable in equilibrium is not affected by the introduction of mandatory disclosure: the positive correlation between trait and quality, which exists when sellers' types are exogenous, disappears when sellers' types are endogenous. Since in the best PEEST with no disclosure the average quality is already at such maximum level, no gain is possible in this regard. As a further consequence, the buyer cannot be better off, since her payoff is increasing in the average quality and, when the average quality is $\overline{\mu}(c_e)$, she is indifferent between any two levels of probability to acquire information. Finally, to see why cost inefficiencies must arise, it is sufficient to observe that there is only one distribution of seller's types that is cost-efficient (abstracting from disclosure costs) and compatible with high quality being produced, namely the one that where only seller's types (H, X) and (L, Y) are present in the market. But this distribution is not sustainable under mandatory labelling with endogenous distribution of seller's types, because there is perfect correlation between traits and qualities, so that the disclosed trait is informative of quality and no buyer would ever pay the disclosure cost to acquire q; this in turn would wipe type (H, X) out of the market.

4 Conclusions

In this paper we have shown that if buyers have the option to acquire information on product quality at a cost, then market unravelling can fail and mandatory disclosure of traits (that a public authority may want to implement as a reaction to the failure of market unravelling) can backfire. In particular, we have demonstrated that mandatory labelling can benefit the buyers (although at the cost of redistributing profits in favor of low quality firms) only if the distribution of quality and labels is given, while it benefits neither buyers nor sellers and necessarily leads to cost inefficiencies if the distribution is determined by competitive pressure on the side of sellers. This result might seem surprising, as the disclosure of relevant information is typically considered to be non-detrimental. However, this need not be the case when buyers can acquire, at a cost, information on their own: the disclosure of relevant information can crowd-out the buyers' incentive to exert effort to acquire information, leading to a problem of asymmetric information that, overall, is more severe.

Future research could explore the case where buyers can also acquire the trait on their own. Intuitively, this could induce (a sort of) market unravelling independently of sellers' behavior, leaving no room for mandatory disclosure (if not for shifting costs from buyers to sellers). However, if the cost to acquire the quality is sufficiently low or if the cost of acquiring the trait is close enough to the cost of acquiring the quality, then we conjecture that our main results still hold: market unravelling can fail because buyers acquire information on quality and, in such case, mandatory disclosure can backfire.

Also, since disclosure costs do not play a crucial role, one can explore the extreme case where voluntary disclosure is costless for sellers. Intuitively, if disclosure does not affect much the behavior of buyers, then high quality sellers with the best trait always disclose (to gain from intuitive buyers) leading to market unravelling. However, this outcome might lead to lower average quality in the long run, for reasons similar to those discussed in this paper. So, it might be of some interest to explore whether a policy aimed at preventing market unravelling (i.e., forbidding labelling) can be desirable in some cases.

Finally, the model in this paper can be extended to include heterogeneity on buyer's cognitive skills, as concerns the cost of information acquisition (with more naive consumers having a higher cost) or the ability to perform correct Bayesian updating (with more naive consumers overweighting priors when elaborating novel information). While our results are likely to be robust to these extensions,[31] it may be interesting to perform a comparative

[31] In a previous version of this paper, we obtain qualitatively analogous results in a model with two buyer's types: deliberative buyers are like buyers in this paper, while intuitive buyers never acquire information due to a very high cost of information acquisition.

statics analysis considering different distributions of cognitive skills.

Let us conclude with a simple remark that tries to answer the following question: if mandatory labelling backfires, what can be done to contrast the problem of asymmetric information? One possibility is to invest on buyers' capabilities, with the aim of reducing the cost of information acquisition. This, reasonably, would entail investments in consumer education and culture, in the collection and diffusion of hard information which are relevant to asses quality, and in the sharing of reliable consumers' feedback.

Acknowledgements

We declare that we have received support from the Italian Ministry of Education, Universities and Research: 2012 PRIN project 2012Z53REX "The Economics of Intuition and Reasoning: a Study On the Change of Rational Attitudes under Two Elaboration Systems (SOCRATES)", and 2017 PRIN project 20178293XT "Cognitive Modes, Social Motives and Prosocial Behavior (COSMOPROBE)".

References

Albert, J. (2010). New technologies and food labelling: the controversy over labelling of foods derived from genetically modified crops. In *Innovations in food labelling*, pp. 153–167. Elsevier.

Anderson, S. P. and R. Renault (2006). Advertising content. *American Economic Review 96*(1), 93–113.

Anderson, S. P. and R. Renault (2013). The advertising mix for a search good. *Management Science 59*(1), 69–83.

Bar-Isaac, H., G. Caruana, and V. Cuñat (2010). Information gathering and marketing. *Journal of Economics & Management Strategy 19*(2), 375–401.

Bar-Isaac, H., G. Caruana, and V. Cuñat (2012). Information gathering externalities for a multi-attribute good. *Journal of Industrial Economics 60*(1), 162–185.

Bertomeu, J. and D. Cianciaruso (2018). Verifiable disclosure. *Economic Theory 65*(4), 1011–1044.

Bilancini, E. and L. Boncinelli (2018a). Rational attitude change by reference cues when information elaboration requires effort. *Journal of Economic Psychology 65*, 90–107.

Bilancini, E. and L. Boncinelli (2018b). Signaling to analogical reasoners who can acquire costly information. *Games and Economic Behavior 110*, 50–57.

Bilancini, E. and L. Boncinelli (2018c). Signaling with costly acquisition of signals. *Journal of Economic Behavior & Organization 145*, 141–150.

Board, O. (2009). Competition and disclosure. *Journal of Industrial Economics 57*(1), 197–213.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and consumer choice. *Journal of Political Economy 121*(5), 803–843.

Brécard, D. (2017). Consumer misperception of eco-labels, green market structure and welfare. *Journal of Regulatory Economics 51*(3), 340–364.

Butler, D. and D. Read (2017). Unravelling & strategic disclosure: Evidence from the hospitality industry.

Celik, L. (2014). Information unraveling revisited: disclosure of horizontal attributes. *Journal of Industrial Economics 62*(1), 113–136.

Chakraborty, A. and R. Harbaugh (2010). Persuasion by cheap talk. *The American Economic Review 100*(5), 2361–2382.

Cheong, I. and J.-Y. Kim (2004). Costly information disclosure in oligopoly. *Journal of Industrial Economics 52*(1), 121–132.

Cho, I.-K. and D. M. Kreps (1987). Signaling games and stable equilibria. *Quarterly Journal of Economics*, 179–221.

Crawford, V. and J. Sobel (1982). Strategic information transmission. *Econometrica*, 1431–1451.

Deversi, M., A. Ispano, and P. Schwardmann (2018). Spin doctors: A model and an experimental investigation of vague disclosure. *CESifo Working Paper No. 7244*.

Dewatripont, M. and J. Tirole (2005). Modes of communication. *Journal of Political Economy 113*(6), 1217–1238.

Dranove, D. and G. Z. Jin (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature 48*(4), 935–63.

Dye, R. A. and S. S. Sridhar (1995). Industry-wide disclosure dynamics. *Journal of Accounting Research*, 157–174.

Emons, W. and C. Fluet (2012). Non-comparative versus comparative advertising of quality. *International Journal of Industrial Organization 30*(4), 352–360.

Evans, J. S. B. and K. E. Stanovich (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science 8*(3), 223–241.

Fishman, M. J. and K. M. Hagerty (2003). Mandatory versus voluntary disclosure in markets with informed and uninformed customers. *Journal of Law, Economics, and Organization 19*(1), 45–63.

Frondel, M., A. Gerster, and C. Vance (2020). The power of mandatory quality disclosure: Evidence from the german housing market. *Journal of the Association of Environmental and Resource Economists 7*(1), 181–208.

Gabaix, X. and D. Laibson (2006). Shrouded attributes, consumer myopia, and information suppression in competitive markets. *Quarterly Journal of Economics 121*(2).

Giovannoni, F. and D. J. Seidmann (2007). Secrecy, two-sided bias and the value of evidence. *Games and Economic Behavior 59*(2), 296–315.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 461–483.

Grossman, S. J. and O. D. Hart (1980). Disclosure laws and takeover bids. *Journal of Finance 35*(2), 323–334.

Guo, L. (2016). Contextual deliberation and preference construction. *Management Science 62*(10), 2977–2993.

Guo, L. and J. Zhang (2012). Consumer deliberation and product line design. *Marketing Science 31*(6), 995–1007.

Hotz, V. J. and M. Xiao (2013). Strategic information disclosure: The case of multiattribute products with heterogeneous consumers. *Economic Inquiry 51*(1), 865–881.

Janssen, M. C. and S. Roy (2014). Competition, disclosure and signalling. *Economic Journal*.

Jin, G. Z., M. Luca, and D. Martin (2020). Is no news (perceived as) bad news? an experimental investigation of information disclosure. *American Economic Journal: Microeconomics (forthcoming).*

Jovanovic, B. (1982). Truthful disclosure of information. *Bell Journal of Economics*, 36–44.

Kahneman, D. (2003). A perspective on judgement and choice. *American Psychologist 58*, 697–720.

Kalaycı, K. and M. Serra-Garcia (2015). Complexity and biases. *Experimental Economics*, 1–20.

Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics 11*, 249–272.

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

Kiesel, K. and S. B. Villas-Boas (2013). Can information costs affect consumer choice? nutritional labels in a supermarket experiment. *International Journal of Industrial Organization 31*(2), 153–163.

Koessler, F. and R. Renault (2012). When does a firm disclose product information? *RAND Journal of Economics 43*(4), 630–649.

Levin, D., J. Peck, and L. Ye (2009). Quality disclosure and competition. *Journal of Industrial Economics 57*(1), 167–196.

Li, S., M. Peitz, and X. Zhao (2014). Information disclosure and consumer awareness. Technical report.

Loewenstein, G., C. R. Sunstein, and R. Golman (2014). Disclosure: Psychology changes everything. *Annual Review of Economics 6*(1), 391–419.

Mas-Colell, A., M. D. Whinston, J. R. Green, et al. (1995). *Microeconomic Theory*, Volume 1. Oxford university press New York.

Matthews, S. and A. Postlewaite (1985). Quality testing and disclosure. *RAND Journal of Economics*, 328–340.

Mengel, F. (2012). On the evolution of coarse categories. *Journal of Theoretical Biology 307*, 117–124.

Milgrom, P. (2008). What the seller won't tell you: Persuasion and disclosure in markets. *Journal of Economic Perspectives 22*(2), 115–131.

Milgrom, P. and J. Roberts (1986). Relying on the information of interested parties. *RAND Journal of Economics*, 18–32.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 380–391.

Okuno-Fujiwara, M., A. Postlewaite, and K. Suzumura (1990). Strategic information revelation. *Review of Economic Studies 57*(1), 25–47.

Schmeiser, S. (2014). Consumer inference and the regulation of consumer information. *International Journal of Industrial Organization 37*, 192–200.

Schmidbauer, E. (2017). Multi-period competitive cheap talk with highly biased experts. *Games and Economic Behavior 102*, 240–254.

Seidmann, D. J. and E. Winter (1997). Strategic information transmission with verifiable messages. *Econometrica*, 163–169.

Shavell, S. (1994). Acquisition and disclosure of information prior to sale. *RAND Journal of Economics*, 20–36.

Stahl, K. and R. Strausz (2017). Certification and market transparency. *Review of Economic Studies 84*(4), 1842–1868.

Sun, M. (2011). Disclosing multiple product attributes. *Journal of Economics & Management Strategy 20*(1), 195–224.

Wang, C. (2013). Advertising as a search deterrent. *Available at SSRN 2404274*.

Appendix: Proofs

Proof of Lemma 1

We start by showing that z^*_μ is continuous in μ. We observe that z^*_μ is identified by the first-order condition

$$\mu\frac{\partial U(z,H)}{\partial z} + (1-\mu)\frac{\partial U(z,L)}{\partial z} = 0, \tag{2}$$

whose derivative with respect to z is

$$\mu\frac{\partial^2 U(z,H)}{\partial z^2} + (1-\mu)\frac{\partial^2 U(z,L)}{\partial z^2}, \tag{3}$$

which is negative (hence different from 0) because $\partial^2 U(z,H)/\partial z^2 < 0$ and $\partial^2 U(z,L)/\partial z^2 < 0$ for all z. Therefore, by the implicit function theorem, we can conclude that z^*_μ is continuous in μ.

Now, since $U(z,q)$ is continuous in z, we have that $\lim_{\mu\to 0}\mu U(z^*_\mu, H) + (1-\mu)U(z^*_\mu, L) = U(z^*_L, L)$. Moreover, it is immediate to observe that $\lim_{\mu\to 0}\mu U(z^*_H, H) + (1-\mu)U(z^*_L, L) = U(z^*_L, L) - c_e$. Since the functions involved are continuous, there exists a threshold $\underline{\mu}(c_e)$ such that, if $\mu \in [0, \underline{\mu}(c_e)]$, then $(\underline{e}, z^*_\mu)$ is optimal for the buyer. Moreover, if $\mu < \underline{\mu}(c_e)$ then the buyer strictly prefers $(\underline{e}, z^*_\mu)$.

Analogously, we also have that $\lim_{\mu\to 1}\mu U(z^*_\mu, H) + (1-\mu)U(z^*_\mu, L) = U(z^*_H, H)$, and that $\lim_{\mu\to 1}\mu U(z^*_H, H) + (1-\mu)U(z^*_L, L) = U(z^*_H, H) - c_e$. Again, since the functions involved are continuous, there exists a threshold $\overline{\mu}(c_e)$ such that, if $\mu > \overline{\mu}(c_e)$, then $(\underline{e}, z^*_\mu)$ is optimal for the buyer. Moreover, if $\mu > \overline{\mu}(c_e)$ then the buyer strictly prefers $(\underline{e}, z^*_\mu)$.

Proof of Lemma 2

The difference between the maximum expected payoff earned under $\overline{e}$ and under $\underline{e}$ can be written as $\mu[U(z^*_H, H) - U(z^*_\mu, H)] + (1-\mu)[U(z^*_L, L) - U(z^*_\mu, L)] - c_e$. If $c_e = 0$, then such an expression is surely positive, because $[U(z^*_H, H) - U(z^*_\mu, H)] > 0$ and $[U(z^*_L, L) - U(z^*_\mu, L)] > 0$. Since the functions involved are continuous, there exists a threshold $\hat{c}_e(\mu)$ such that, if $c_e \le \hat{c}_e(\mu)$, then $(\overline{e}, z^*_L, z^*_H)$ is an optimal response for the buyer. Moreover, if $c_e < \hat{c}_e(\mu)$, then the buyer strictly prefers $(\overline{e}, z^*_L, z^*_H)$.

Proof of Proposition 1

From Lemmas 1 and 2 we know that there exist $\underline{\mu}(c_e)$ and $\overline{\mu}(c_e)$ such that $(\underline{e}, z^*_\mu)$ is optimal for $\mu \in [0, \underline{\mu}(c_e)] \cup [\overline{\mu}(c_e), 1]$. Here we show that, if there exists two distinct beliefs μ' and

μ'' where $(\overline{e}, z_L^*, z_H^*)$ is optimal then it is a fortiori optimal for any convex combination of μ' and μ''. These results imply that the set of beliefs for which $(\overline{e}, z_L^*, z_H^*)$ is optimal is either empty or an interval.

This would guarantee that there exist at most three intervals. To this aim, we prove the strict concavity of the difference between the expected utilities of choosing $\overline{e}$ and $\underline{e}$.

Without loss of generality we assume $\mu' < \mu''$. We now compute such a difference of expected utilities for both μ' and μ'':

$$\mu'[U(z_H^*, H) - U(z_{\mu'}^*, H)] + (1-\mu')[U(z_L^*, L) - U(z_{\mu'}^*, L)] - c; \tag{4}$$

$$\mu''[U(z_H^*, H) - U(z_{\mu''}^*, H)] + (1-\mu'')[U(z_L^*, L) - U(z_{\mu''}^*, L)] - c. \tag{5}$$

We now consider a convex combination of μ' and μ'':

$$\hat{\mu} = \alpha\mu' + (1-\alpha)\mu'', \tag{6}$$

and we compute the same difference of expected utilities also for $\hat{\mu}$:

$$\hat{\mu}[U(z_H^*, H) - U(z_{\hat{\mu}}^*, H)] + (1-\hat{\mu})[U(z_L^*, L) - U(z_{\hat{\mu}}^*, L)] - c. \tag{7}$$

We now take the difference between (7) and the convex combination of (4) and (5) with weights α and $1-\alpha$, obtaining after a few simplifications the following expression:

$$\begin{aligned}&\alpha\left[\mu' U(z_{\mu'}^*, H) + (1-\mu')U(z_{\mu'}^*, L) - \mu' U(z_{\hat{\mu}}^*, H) - (1-\mu')(U(z_{\hat{\mu}}^*, L))\right] + \\ &+(1-\alpha)\left[\mu'' U(z_{\mu''}^*, H) + (1-\mu'')U(z_{\mu''}^*, L) - \mu'' U(z_{\hat{\mu}}^*, H) - (1-\mu'')(U(z_{\hat{\mu}}^*, L))\right].\end{aligned} \tag{8}$$

We conclude by observing that the two terms in square brackets are both positive, because $z_{\mu'}^*$ is optimal against belief μ', $z_{\mu''}^*$ is optimal against belief μ'', and z_μ^* is strictly increasing in μ, so that $z_{\mu'}^* \neq z_{\hat{\mu}}^* \neq z_{\mu''}^*$; hence, (8) is positive, which means that the difference between the expected utilities of choosing $\overline{e}$ and $\underline{e}$ is strictly concave.

As a final remark, we observe that the difference between the expected utilities of choosing $\overline{e}$ and $\underline{e}$ has a maximum in $[0, 1]$ by the extreme value theorem, because we are considering the difference of two continuous functions. Such a maximum value can be written as $M - c$. We conclude by setting $\hat{c} = M$.

Proof of Proposition 2

If no trait is observed, then by Bayes rule we have that B must hold a belief $\mu = p(H)$. By Lemma 2 we know that, if $c_e < \hat{c}_e(p(H))$ then $(\overline{e}, z_L^*, z_H^*)$ is the buyer's optimal response when belief is $p(H)$.

Given this response, sellers of types (H, X) and (H, Y) who do not disclose t obtain a payoff equal to $V(z_H^*)$, which is clearly larger than anything they can obtain by disclosing t, while sellers of types (L, X) and (L, Y) who do not disclose t obtain $V(z_L^*)$, which may be larger than what they obtain by disclosing t depending on buyer's out-of-equilibrium beliefs when t is observed. Clearly, this is the case for pessimistic beliefs, because L-types do not obtain any gain by disclosing t while still paying the disclosure cost. Hence the profile under consideration, together with pessimistic beliefs, is a wPBE. In the following we show that pessimistic beliefs also pass the requirements imposed by the D1 criterion.

Let $T_H \subset [0, 1]$ be the set of beliefs that entail an optimal response by B such that seller's types (H, X) or (H, Y) obtain a payoff which is least as large as $V(z_H^*)$. Similarly, let $T_L \subset [0, 1]$ be the set of beliefs that entail an optimal response by B such that seller's types (L, X) or (L, Y) obtain a payoff which is least as large as $V(z_L^*)$. As already argued, high quality sellers can never gain by disclosing t, hence $T_H = \emptyset$. The D1 criterion imposes that $\mu(t) = 0$, $t = X, Y$, if $T_H \subset T_L$, and does not impose anything if $T_H = T_L$. Since pessimistic beliefs require that $\mu(t) = 0$, in any case they pass the requirements imposed by the D1 criterion.

Proof of Proposition 3

When $c_e < \hat{c}_e(p(H))$, by Proposition 2 the pooling equilibrium (σ^p, β^p) exists. In (σ^p, β^p) the sellers of types (H, t), $t = X, Y$, are recognized as high quality, hence their payoff is equal to $V(z_H^*)$. In the mandatory labelling equilibrium $(\sigma^m, \beta^m$, the buyer's expenditure cannot be larger than z_H^*, and the seller pays the disclosure cost c_d. This gives the result in (i).

Consider the seller of type (L, X). Since in (σ^p, β^p) the type (L, X) is recognized as low quality, she obtains a payoff equal to $V(z_L^*)$. In the mandatory labelling equilibrium (σ^m, β^m), if $p(H|X) > \overline{\mu}(c_e)$, then the buyer does not acquire q, and hence the type (L, X) obtains a payoff equal to $V(z_{P(H|X)}^*) - c_d$. So, in this case type (L, X) gains from mandatory labelling if $V(z_{P(H|X)}^*) - c_d > V(z_L^*)$, which gives the result in (ii).

Consider the seller of type (L, Y). The result in (iii) follows from the same argument used for type (L, X).

Proof of Proposition 4

When $c_e < \hat{c}_e(p(H))$, by Proposition 2 the pooling equilibrium (σ^p, β^p) exists. In (σ^p, β^p) the buyer's expected payoff is equal to:

$$p(H)U(z_H^*, H) + (1 - p(H))U(z_L^*, L) - c_e. \tag{9}$$

In the mandatory labelling equilibrium (σ^m, β^m) her expected payoff is also equal to (9) if $p(H|X) < \overline{\mu}_{c_e}$ and $p(H|Y) > \underline{\mu}_{c_e}$, because $(\overline{e}, z_L^*, z_H^*)$ is still an optimal response. Hence, mandatory labelling provides no benefit to her in such case, which gives the result in (i).

Suppose now that either $p(H|X) < \overline{\mu}_{c_e}$, or $p(H|Y) > \overline{\mu}_{c_e}$, or both. In such cases $(\underline{e}, z_{\mu(t)}^*)$ is the only optimal response for at least one observed trait; this means that it pays more than $(\overline{e}, z_L^*, z_H^*)$ thus making the buyer's expected payoff larger, which gives the result in (ii).

Proof of Proposition 5

Firstly, we show that no PEEST with $P(H) > \overline{\mu}(c_e)$ can exist, so that if a PEEST exists with $p(H) = \overline{\mu}(c_e)$ it has the highest $p(H)$. Consider a PEEST such that $p(H) > \overline{\mu}(c_e)$; then, we must have that $p(H|X) > \overline{\mu}(c_e)$ or $p(H|Y) > \overline{\mu}(c_e)$ or both. If $p(H|t) > \overline{\mu}(c_e)$ for some $t \in \{X, Y\}$ then, by Lemma 1 and Bayes rule, the buyer must optimally respond choosing $\underline{e}$ and $z_{p(H|t)}^*$. So, if t is disclosed, then type (L, t) makes greater profits than type (H, t) which, by condition (ii), implies that $p(H|t) = 0$. Since this holds for all t, $p(H) > \overline{\mu}(c_e)$ is impossible if t is disclosed. Suppose t is not disclosed. Then, by Lemma 1 and Bayes rule, the buyer must optimally respond choosing $\underline{e}$ and $z_{p(H)}^*$. This in turn implies that L-types always make greater profits than H-types which, by condition (ii), implies that $p(H) = 0$.

Now, we show that $(\sigma^p, \tilde{\beta}^p)$ satisfies condition (i) of the definition of PEEST for $p(H) = \overline{\mu}(c_e)$, $p(H|X) = 1$ and $p(H|Y) = 0$. In particular, we first deduce the optimal response by the buyer given σ^p and $p(H) = \overline{\mu}(c_e)$, and from this we show that the set of beliefs T_H, which justifies a deviation by a H-type, is either empty or strictly contained in the set T_L, which justifies a deviation by a L-type.

From σ^p it follows that no trait is observed. By Bayes rule, the buyer's belief conditional to not observing any label must be equal to $\overline{\mu}(c_e)$. Therefore, by means of Lemmas 1 and 2 it is straightforward to conclude that the buyer is indifferent between $(\overline{e}, z_L^*, z_H^*)$ and $(\underline{e}, z_{\overline{\mu}(c_e)}^*)$. Hence, any acquisition probability in $[0, 1]$ is a best response when no trait is observed, provided that the expenditure choices are: z_H^* if $q = H$ is observed, z_L^* if $q = L$ is observed, and $z_{\overline{\mu}(c_e)}^*$ if quality remains unobserved.

Given this response, sellers of types (H, t) who do not disclose t obtain a payoff equal to $\lambda^* V(z_H^*) + (1 - \lambda^*)V(z_{\overline{\mu}(c_e)}^*)$, gross of cost c_{Ht}. If they instead disclose t, they can obtain at most a payoff equal to: $V(z_H^*) - c_d$ if $\mu \in [\underline{\mu}(c_e), \overline{\mu}(c_e)]$ with the buyer optimally choosing $(\overline{e}, z_L^*, z_H^*)$ (by Lemma 2), and equal to $V(z_\mu^*) - c_d$ if $\mu \in [0, \underline{\mu}(c_e)] \cup [\overline{\mu}(c_e), 1]$ with the buyer optimally choosing $(\underline{e}, z_\mu^*)$ (by Lemma 1). Consider the case of $\mu \in [\underline{\mu}(c_e), \overline{\mu}(c_e)]$ with the buyer optimally choosing $(\overline{e}, z_L^*, z_H^*)$. Then, not disclosing is surely strictly better than disclosing if:

$$V(z_H^*) - V(z_{\overline{\mu}(c_e)}^*) < \frac{c_d}{1-\lambda^*}. \tag{10}$$

Since $\overline{\mu}(c_e)$ tends to 1 as c_e tends to 0, there always exists $\check{c}_e > 0$ such that the two sides of (10) are equal to each other. Therefore, for all $c_e < \check{c}_e$ there is no $\mu \in (\underline{\mu}(c_e), \overline{\mu}(c_e))$ that can justify the choice of types (H, t) to disclose t. Now, we consider the case of $\mu \in [0, \underline{\mu}(c_e)] \cup [\overline{\mu}(c_e), 1]$ with the buyer optimally choosing $(\underline{e}, z_\mu^*)$. For such beliefs, not disclosing is strictly better than disclosing if:

$$\begin{aligned} \lambda^* V(z_H^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*) \quad &> \quad V(z_\mu^*) - c_d \quad \Leftrightarrow \\ V(z_\mu^*) - [\lambda^* V(z_H^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*)] \quad &< \quad c_d. \end{aligned} \tag{11}$$

Given the buyer's response, sellers of types (L, t) who do not disclose t obtain $\lambda^* V(z_L^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*)$, gross of cost c_{Lt}. Let $T_L \subset [0,1]$ be the of set of beliefs that, if held by B, entail an optimal response by B such that seller's types (L, X) or (L, Y) obtain a payoff which is larger than $\lambda^* V(z_L^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*)$. Since $\lambda^* V(z_H^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*)$ is not smaller than $\lambda^* V(z_L^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*)$, if μ is such that (11) is satisfied then it is also such that $V(z_\mu^*) - c_d > V\lambda^* V(z_L^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*)$, implying that $T_H \subseteq T_L$. The D1 criterion imposes that $\mu(t) = 0$, $t = X, Y$, if $T_H \subset T_L$, and does not impose anything if $T_H = T_L$. Since pessimistic beliefs require that $\mu(t) = 0$, in any case they pass the requirements imposed by the D1 criterion.

To show that $(\overline{\mu}(c_e), (1,0), (\sigma^p, \tilde{\beta}^p))$ satisfies condition (ii) of the definition of PEEST, it is enough to observe that λ^* solves:

$$\begin{aligned} \pi_{HX} = \lambda^* V(z_H^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e))}^*) - c_{HX} = \lambda^* V(z_L^*) + (1-\lambda^*)V(z_{\overline{\mu}(c_e)}^*) - c_{LY} = \pi_{LY} \;\Leftrightarrow \\ \Leftrightarrow \; \lambda^* V(z_H^*) - c_{HX} = \lambda^* V(z_L^*) - c_{LY} \end{aligned} \tag{12}$$

and that, because of $c_{HY} > c_{HX}$ and $c_{LX} > c_{LY}$, both π_{HY} and π_{LX} are strictly lower than $\pi_{HX} = \pi_{LY}$.

Proof of Proposition 6

We observe that under mandatory labelling a PEEST with $p(H|X) > \overline{\mu}(c_e)$ or $p(H|Y) > \overline{\mu}(c_e)$, and hence with $p(H) > \overline{\mu}(c_e)$, can not exist. This is evident if one applies the argument provided in the first paragraph of the proof of Proposition 5 to the case of $\sigma = \sigma^m$, i.e., abstracting from cases where the buyer does not observe any label.

To check that the buyer's payoff cannot increase under mandatory labelling, it suffices to note that such payoff is given by an average of the payoff obtained when x is observed and the payoff obtained when y is observed. Each of these two payoffs, and also the payoff in case no trait is observed, has the same functional expression which is strictly increasing in the average quality (conditional on information being x, or y, or 0). Moreover, when the average quality is $\overline{\mu}(c_e)$, the buyer is indifferent between acquiring the information on quality and not acquiring it, hence also between any two levels of the probability λ of acquiring the information.

Finally, since $c_{HX} < c_{HY}$ and $c_{LX} > c_{LY}$, cost efficiency requires that, if both quality H and quality L are produced, $(p(H|X), p(H|Y)) = (1, 0)$, which is impossible: indeed, by Lemma (1) we know that the buyer would never acquire q if $p(H|X) = 1$, but then the seller of type (L, X) would obtain higher profits than the seller of type (H, X).