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

Ennio Bilancini* Leonardo Boncinelli†

October 11, 2016

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 sophisticated buyers can acquire information on the actual 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 actual quality of the product. Our main finding is that market unravelling can fail because of the presence of many sophisticated buyers, 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 naive buyers gain while profits decrease for high quality sellers. Further, if the label is sufficiently informative, then also sophisticated buyers gain and profits increase for low quality sellers. When, instead, the joint distribution of qualities and traits is endogenous, mandatory labelling can not lead to an increase in quality or in buyer's utility. Moreover, if average quality is left unaltered, then sellers' profits are not increased and cost-inefficiencies arise.

JEL classification code: D82, D83, L15.

Keywords: certification, cognitive resources, costly acquisition of information, labelling, mandatory disclosure, persuasion, unraveling.

*Dipartimento di Economia "Marco Biagi", Università degli Studi di Modena e Reggio Emilia, Viale Berengario 51, 43 ovest, 41121 Modena, Italia. Tel.: +39 059 205 6843, fax: +39 059 205 6947, email: ennio.bilancini@unimore.it.

†Dipartimento di Scienze per l'Economia e l'Impresa, Università degli Studi di Firenze, Via delle Pandette 9, 50127 Firenze, Italia. Tel.: +39 055 2759578, fax: +39 055 2759910, email: leonardo.boncinelli@unifi.it.

1 Introduction

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. In particular, in many cases buyers have the option to acquire information on the actual 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, typically for sellers it is not feasible to disclose (or the government can not impose the disclosure of) exact information on quality, while 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 this paper we develop a variant of the persuasion game by Milgrom and Roberts (1986) that takes into account these features. In our model, a seller can be of high or low quality and has one of two traits that correlate differently with quality (one more than the other). Initially, both quality and trait are private information of the seller. The buyer can be a sophisticated type, who is able to learn the seller's quality by incurring an acquisition cost, or a naive type, who never learns actual quality (e.g., because the acquisition cost is always too high for him). In any case, before sophisticated 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. 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 first conduct the analysis assuming that the joint distribution of quality and traits is exogenous, and then we extend the analysis endogenizing the distribution.

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 either the cost to acquire information for sophisticated buyers is quite large or naive buyers are too many. On the contrary, whenever the cost to acquire information is small and the number of sophisticated buyers is large, 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 belive that a buyer 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. More precisely, 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 sophisticated buyers – if it allows them to save on the acquisition cost – and always benefits naive buyers – as they get more informed – but it can backfire on the seller's side. In particular, profits can increase for low quality sellers, and always decrease for high quality sellers. This happens whenever the observation of the label – which helps buyers to have more precise beliefs about product quality – crowds out the incentive to exert the costly effort which allows sophisticated buyers to learn the actual quality of the sold product. 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.[1]

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, to be 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 any reasonable profit-monotone dynamics. We show that, whenever the cost to acquire information is small

[1] Schmeiser (2014) studies mandatory disclosure on a different piece of information: the relative importance of product attributes. In such setup, he shows that inferential mistakes can lead to over or under-regulation by regulators and over or under-estimatimation of the importance of product attributes by consumers.

enough and the number of sophisticated buyers is large 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 effort of sophisticated buyers to learn quality. Labelling provides buyers with some extra information on quality. For naive buyers, this is good. Also for sophisticated buyers this is good. However, for the latter it also creates an incentive not to acquire more precise information on quality, since the acquisition is costly. But if sophisticated buyers 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 sophisticated 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 obtain 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 one trait and low quality sellers specialize in the other trait, even when this is cost efficient.

The paper is organized as follows. In Section 2 we review the related literature, indicating in what respects our model is different from existing models. In Section 3 we describe the persuasion game with labelling, assuming an exogenous joint distribution of qualities and traits. In Section 4 we analyze the model and characterize the pooling equilibrium with no disclosure, showing that it survives D1. In Section 5 we study the effects of mandatory labelling starting from the pooling equilibrium of Section 4. In Section 6 we study a variant of the model of Section 3 where the joint distribution of qualities and traits is endogenous; we characterize the equilibrium with highest quality for which no voluntary disclosure emerges and we assess the effects of mandatory disclosure starting from it. Section 7 discusses some of the crucial assumptions of our model and then concludes.

2 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). 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).[2] 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. [3]

The second stream of literature is more focused on market disclosure and on certification costs (see Dranove and Jin, 2010, for a recent 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.

[2] 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.

[3] Other variants of persuasion games have been studied. Anderson and Renault (2013), bulding 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 outcome is a separating equilibrium with quality unravelling, but they also show that unravelling may fail for low enough search costs.

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.[4] 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. 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.[5] 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 leads the buyers to believe that they are in front of a low quality seller, because the fact that sophisticated buyers acquire information on their own 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 quality sellers.

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). 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

[4] 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.

[5] 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.

behavior. Our model follows this literature on limited cognitive resources by assuming that buyers incur a cost to process available information on quality, but we maintain the standard economic assumption that the buyer is fully aware of this issue and optimally allocates her cognitive resources. In particular, we follow Dewatripont and Tirole (2005) and, more closely, Bilancini and Boncinelli (2014) by interpreting the choice of the sophisticated buyer to acquire information on quality 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. This distinguishes our approach from models as in (Fishman and Hagerty, 2003) where buyers can never choose to acquire more information at a cost.[6]

3 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. Furthermore, there are two types of B, one naive denoted by n and one sophisticated denoted by s. The prior probability $p(b)$ that B's types is $b \in \{n, s\}$ is

[6] 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.

given by $p(n)$ and $p(s) = 1 - p(n)$. (To fix ideas, $p(n)$ can be interpreted as the fraction of buyers in the population who are naive.) The sophisticated buyer can exert effort and incur the cost $c_e > 0$ to learn q. The naive buyer can not acquire q on her own – i.e., the cost to acquire q is too large for type n. So, naive buyers may observe t (if disclosed by some type of S), while sophisticated buyers may observe t, q, or both. In any case, after that trait and/or quality have had the chance to be observed, each buyer type chooses some scalar z, which can be interpreted as the amount of money that she spends on the firm's products.[7]

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.[8] 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. 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 $\sigma : \{H, L\} \times \{X, Y\} \to \{0, 1\}$ 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 pair of functions, one for the naive type and one for the sophisticated type. For type n the function is $\beta_n : \{0, x, y\} \to \{s_1\} \times \mathbb{R}_+$,while for type s the function is the function $\beta_s : \{0, x, y\} \to \{s_1\} \times \mathbb{R}_+ \cup \{s_2\} \times \mathbb{R}_+^2$ 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 effort to acquire q, denoted by s_2, or not, denoted by s_1, and how much to spend, denoted by $z \in \mathbb{R}_+$.[9] In case s_2 is chosen, an additional information is acquired by B regarding

[7] As suggested by Milgrom (2008), the scalar z can be interpreted as, e.g., the quantity that the buyer purchases or the highest price that she is willing to pay for a unit. Also, z can be intepreted interpreted as a reduced form of a model where the seller is choosing its price optimally, anticipating what the buyers purchase behavior is going to be.

[8] 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.

[9] This labelling owes to the classification of elaboration processes as "System 1", or S1, which is fast, cheap and intuitive, and "System 2", or S2, which 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.

whether the quality is H or L, hence in such a case the buyer actually chooses a pair of numbers, $z_L \in \mathbb{R}_+$ and $z_H \in \mathbb{R}_+$, the first in case L is discovered and the second in case H is discovered.

As a solution concept we focus on the *weak Perfect Bayesian Equilibrium* (wPBE) in pure strategies. 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 make use 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.

4 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 that the product is of high quality. If B is of type n or if B is of type s but chooses s_1 (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 is of type s and chooses s_2 (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 behavior of the sophisticated buyer type, we have to compare $\mu U(z^*_\mu, H) + (1-\mu)U(z^*_\mu, L)$, which is the maximum expected payoff earned if s_1 is chosen, and $\mu U(z^*_H, H) + (1-\mu)U(z^*_L, L)$, which is the maximum expected payoff earned if s_2 is chosen. This leads us to the following propositions:

PROPOSITION **1.** *(Optimality of s_1 for type s)*
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 (s_1, z^*_μ) is an optimal response for the sophisticated buyer. In addition, if $\mu \in [0, \underline{\mu}(c_e)) \cup (\overline{\mu}(c_e), 1]$ it is the only optimal response.

PROPOSITION **2.** *(Optimality of s_2 for type s)*
For any given belief on quality μ, there exists $\hat{c}_e(\mu)$ such that, if $c_e \leq \hat{c}_e(\mu)$, then (s_2, z^*_L, z^*_H)

is an optimal response for the sophisticated buyer. In addition, if $c_e < \hat{c}_e(\mu)$ *it is the only optimal response.*

Propositions 1 and 2 allow us to conclude that, if the acquisition cost c_e is sufficiently low, then the optimal choice of the sophisticated buyer as a function of μ can be summarized graphically as in Figure 1. 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 sophisticated buyer not to exert effort 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, and is equal to the optimal amount chosen by the naive buyer. However, when the belief is intermediate – i.e., $\mu \in [\underline{\mu}(c_e), \overline{\mu}(c_e)]$ – the sophisticated buyer finds it optimal to pay the acquisition cost c_e and acquire the exact knowledge of quality. In this range of beliefs the optimal amount z^* will be conditioned to the quality discovered, 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. In this same range of beliefs the choice of the naive buyer differs from the choice of the sophisticated buyer in that it is both unconditional on q and increasing in μ.

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. Also, the probability $1 - p(n)$ that the buyer is sophisticated matters, as this affects the likelihood that B acquires q. 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 sophisticated buyer chooses to acquire q she also chooses z conditionally on $q = H$ and $q = L$ (see Figure 1). In particular, when c_d is not too large, type (H, X) can gain from disclosing the trait x even if the sophisticated buyer acquires q, provided that $p(n)$ is large enough. More precisely, a necessary condition for type (H, X) to profit by disclosing x when $p(H) \in (\underline{\mu}(c_e), \overline{\mu}(c_e))$ is that $c_d < p(n)[(V(z^*_H) - V(z^*_{p(H)})]$, i.e., that the maximal increase in expenditure by naive buyers is greater than the cost of disclosing. This condition gives the following upper bound on $p(n)$:

$$\hat{p}(c_d, p(H)) = \frac{c_d}{V(z^*_H) - V(z^*_{p(H)})} \tag{1}$$

that is non-negative and increasing in c_d and $p(H)$. In other words, when $p(n) \leq \hat{p}(c_d, p(H))$, if the sophisticated buyer acquires q and behaves optimally then type (H, X) maximizes his

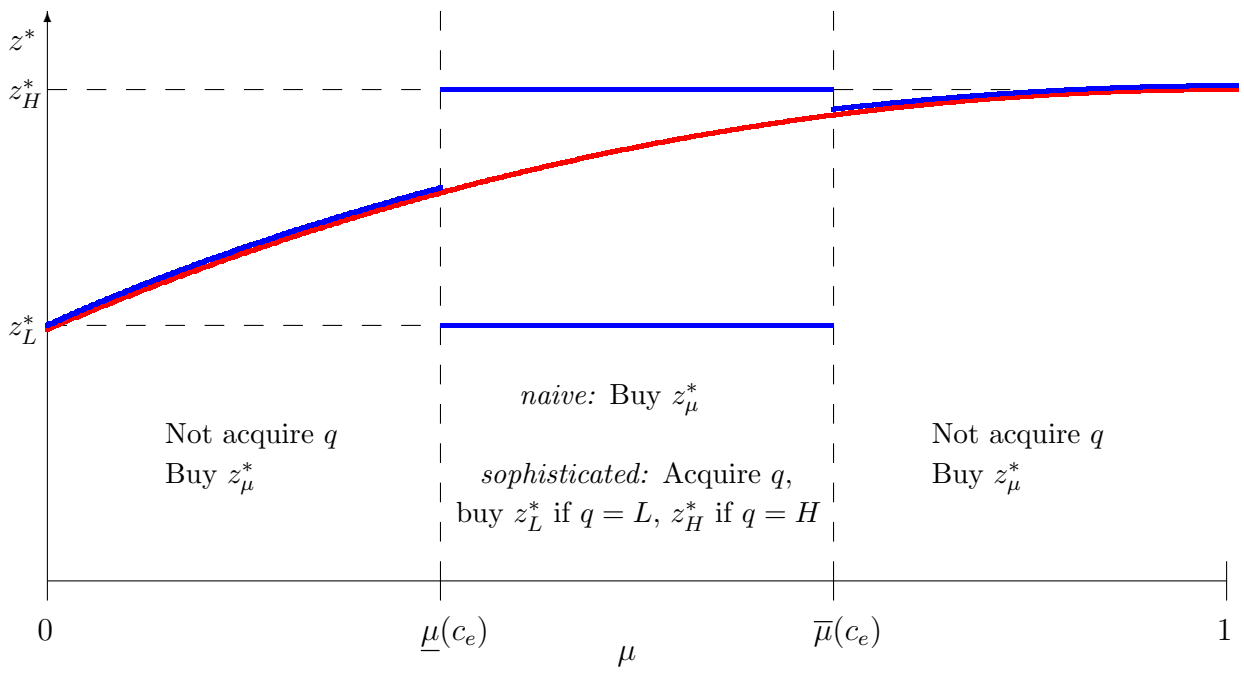

Figure 1: Buyer's optimal behavior as a function of beliefs μ on quality. The continuous (red) map represents the optimal choice for the naive buyer, while the piecewise (blue) map represents the optimal choice for the sophisticated buyer.

payoff by not disclosing x. We observe that, a fortiori, this also holds for type (H, Y) who gains strictly less by disclosing y, while it does not hold for type (L, X) who can also gain from inducing the sophisticated buyer to switch from acquiring q to not acquiring it.

We observe 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 **3.** *(Pooling equilibrium with no voluntary disclosure)*
For any given prior belief $p(H)$, if $c_e < \hat{c}_e(p(H))$ and $p(n) < \hat{p}(c_d, p(H))$, then there exists a wPBE where no seller type discloses, the naive buyer chooses $z^*_{p(H)}$, and the sophisticated buyer acquires q and chooses z_H^* if $q = H$ and z_L^* if $q = L$. This equilibrium survives the D1 criterion.

The intuition underlying Proposition 3 is straightforward. For $c_e < \hat{c}_e(p(H))$, sophisticated

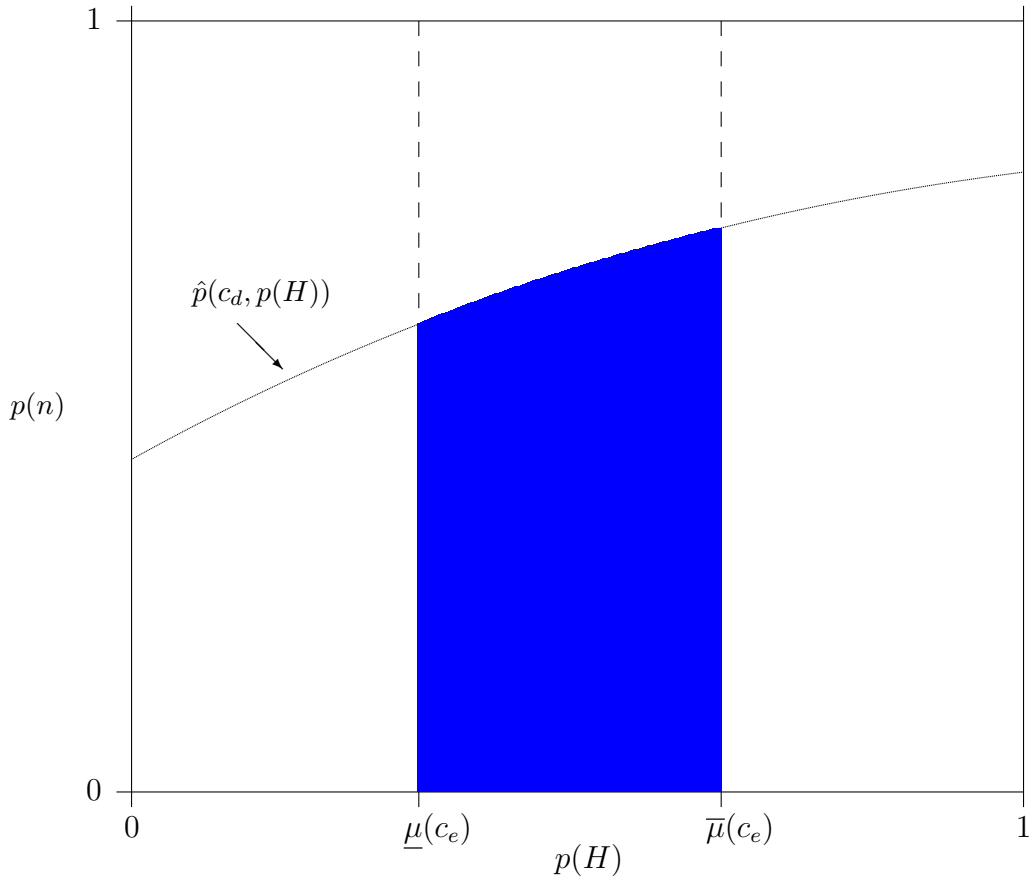

Figure 2: The depicted region (shaded blue) in the space of priors represents the values of $p(H)$ and $p(n)$ for which there exists a pooling equilibrium where no seller type discloses, the naive buyer chooses $z^*_{p(H)}$, and the sophisticated buyer acquires q and chooses z^*_H if $q = H$ and z^*_L if $q = L$.

buyers prefer to acquire q and spend conditionally on it. Given this, for $p(n) < \hat{p}(c_d, p(H))$ type (H, X) maximizes his profits by not disclosing q, independently of buyers' beliefs on quality. A fortiori, this holds for (H, Y). If also L-types do not gain by disclosing, then there is indeed a wPBE with no disclosure (and D1 has no bite on out of equilibrium beliefs). If, instead, L-types gain by disclosing for buyers' beliefs sufficiently high, then D1 requires that a buyer observing any t believes that $q = L$, with the result that no seller's type can gain by disclosing and, hence, this is a wPBE with no disclosure (robust to D1).

Type	**Effect due to naive buyers**	**Effect due to sophisticated buyers**		**Total effect net of c_d**
		if: $p(H\|X) > \overline{\mu}(c_e)$	**if:** $p(H\|Y) < \underline{\mu}(c_e)$	
seller (H,X)	$p(n)[(V(z^*_{p(H\|x)}) - V(z^*_{p(H)})]$ **(+)**	$p(s)[(V(z^*_H) - V(z^*_{p(H\|x)})]$ **(−)**	nil **(0)**	**(−)**
seller (H,Y)	$p(n)[(V(z^*_{p(H\|y)}) - V(z^*_{p(H)})]$ **(−)**	nil **(0)**	$p(s)[(V(z^*_H) - V(z^*_{p(H\|y)})]$ **(−)**	**(−)**
seller (L,X)	$p(n)[(V(z^*_{p(H\|x)}) - V(z^*_{p(H)})]$ **(+)**	$p(s)[(V(z^*_L) - V(z^*_{p(H\|x)})]$ **(+)**	nil **(0)**	**(?)**
seller (L,Y)	$p(n)[(V(z^*_{p(H\|y)}) - V(z^*_{p(H)})]$ **(−)**	nil **(0)**	$p(s)[(V(z^*_L) - V(z^*_{p(H\|y)})]$ **(+)**	**(?)**

Table 1: Effects of mandatory labelling on seller's types.

Let us conclude the section with one remark on what happens when the conditions $c_e < \hat{c}_e(p(H))$ and $p(n) < \hat{p}(c_d, p(H))$ are not jointly satisfied. For $c_e \geq \hat{c}_e(p(H))$ equilibria are possible where X-types voluntarily disclose their trait. This happens if sophisticated sellers do 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. Also for $p(n) \geq \hat{p}(c_d, p(H))$ equilibria with voluntary disclosure of x are possible, but in this case the reason is that the mass of sophisticated types is too little to sustain the profits of H-types, so that, even if all sophisticated buyers acquire q, type (H,X) can find it profitable to disclose x, inducing again type (L,X) to do the same. 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 3 guaratees that they exist for reasonable parameter values. To put it differently, our analysis begins with the following observation: if there are enough sophisticated buyers and they are sufficiently sophisticated (in the sense that their c_e is 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).

Type	Effect for any $p(H\|t)$	Effect if: $p(H\|X) > \overline{\mu}(c_e)$	Effect if: $p(H\|Y) < \underline{\mu}(c_e)$	Total effect
buyer n	(+)	0	0	(+)
buyer s	0	(+)	(+)	(+)

Table 2: Effects of mandatory labelling on buyer's types.

5 Mandatory labelling

Let us denote with $(\sigma^p, \beta_n^p, \beta_s^p)$ an equilibrium profile that leads to a pooling outcome as described in Proposition 3. Also, let us denote with $(\sigma^m, \beta_n^m, \beta_s^m)$ 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.[10] 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 naive buyer chooses β_n^m depending on her priors, i.e., $\beta_n^m(x) = (s_1, z^*_{p(H|X)})$ and $\beta_n^m(y) = (s_1, z^*_{p(H|Y)})$, while the sophisticated buyer chooses β_s^m as follows: $\beta_s^m(x) = (s_2, z_H^*, z_L^*)$ if $p(H|X) < \overline{\mu}(c_e)$, $\beta_s^m(x) = (s_1, z^*_{p(H|X)})$ if $p(H|X) > \overline{\mu}(c_e)$ (being indifferent between s_1 and s_2 if $p(H|X) = \overline{\mu}(c_e)$), and $\beta_s^m(y) = (s_2, z_H^*, z_L^*)$ if $p(H|X) > \underline{\mu}(c_e)$, $\beta^m(y) = (s_1, z^*_{p(H|Y)})$ if $p(H|y) < \underline{\mu}(c_e)$ (being indifferent between s_1 and s_2 if $p(H|y) = \underline{\mu}(c_e)$).

By comparing $(\sigma^p, \beta_n^p, \beta_s^p)$ with $(\sigma^m, \beta_n^m, \beta_s^m)$, we can analyze the consequences of mandatory labelling. We first consider H types. One straightforward effect is that they must incur the cost c_d.[11] Further, since in $(\sigma^p, \beta_n^p, \beta_s^p)$ the high types are selling z_H^* to the sophisticated buyer and $z^*_{p(H|t)}$ to the naive buyer, there are other potential effects: the sophisticated buyer can be induced to use s_1, which reduces her spending from z_H^* to $z^*_{p(H|t)}$, $t = X, Y$, while the naive buyer increases her spending for type (H, X) from $z^*_{p(H)}$ to $z^*_{p(H|X)}$, and reduces her spending for type (H, Y) from $z^*_{p(H)}$ to $z^*_{p(H|Y)}$. The net effect turns out to be always negative for H types. Intuitively, since in the original pooling equilibrium H types could have disclosed t by themselves but preferred not to do so, mandatory disclosure can not be beneficial to them. The following proposition makes this claim precise:

PROPOSITION 4. *(Effects of mandatory labelling on types (H, X) and (H, Y))*

[10] 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.

[11] 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.

Let $c_e < \hat{c}_e(p(H))$ and $p(n) < \hat{p}(c_d, p(H))$. Starting from a pooling equilibrium $(\sigma^p, \beta_n^p, \beta_s^p)$, the introduction of mandatory labelling leads to a mandatory equilibrium $(\sigma^m, \beta_n^m, \beta_s^m)$ where the payoff earned by the seller of type (H,t), $t = X, Y$, is lower.

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, and this is symmetrical for type (L,X) and type (L,Y). Further, there is an effect that is positive for the type (L,X) and negative for the type (L,Y): the naive buyer modifies her spending from $z^*_{p(H)}$ for all seller's types to $z^*_{p(H|X)} > z^*_{p(H)}$ for type (L,X) and to $z^*_{p(H|Y)} < z^*_{p(H)}$ for type (L,Y). Finally, 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 sophisticated buyer to use s_1, then type (L,X) increases sales from z^*_L to $z^*_{p(H|X)}$ for this buyer. A similar effect can arise for type (L,Y) if the buyer's prior conditionally on y is sufficiently low, but the increase is z^* is obviously lower than for type (L,X). So, on balance, type (L,X) is more likely to gain from mandatory disclosure than type (L,Y). Table 1 illustrates all these effects. The following proposition summarizes:

PROPOSITION **5.** *(Effects of mandatory labelling on types (L,X) and (L,Y))*
Let $c_e < \hat{c}_e(p(H))$ and $p(n) < \hat{p}(c_d, p(H))$. Starting from a pooling equilibrium $(\sigma^p, \beta_n^p, \beta_s^p)$, the introduction of mandatory labelling leads to a mandatory equilibrium $(\sigma^m, \beta_n^m, \beta_s^m)$ where:

(i) the payoff earned by the seller of type (L,X) increases if $p(H|X) > \overline{\mu}(c_e)$ and $c_d < (1-p(n))[V(z^*_{p(H|X)}) - V(z^*_L)] + p(n)[(V(z^*_{p(H|X)}) - V(z^*_{p(H)})]$,
(ii) the payoff earned by the seller of type (L,Y) increases if $p(H|y) < \underline{\mu}(c_e)$ and $c_d < (1-p(n))[V(z^*_{p(H|Y)}) - V(z^*_L)] + p(n)[(V(z^*_{p(H|Y)}) - V(z^*_{p(H)})]$.

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 one positive effect due to the additional information conveyed by t which is always present but that, if priors $p(H|t)$ are very close to $p(H)$, is experienced only by the naive buyer. Indeed, if priors $p(H|t)$ are very close to $p(H)$, then the observation of t leaves the sophisticated buyer still too uncertain about quality, so that she prefers to keep using s_2 (paying c_e and acquiring q). This positive effect for naive buyers is increasing in the distance $|p(H) - p(H|t)|$, $t = X, Y$, i.e., it is increasing in the amount of information disclosed by the label. Further, when priors $p(H|t)$ are sufficiently distant from $p(H)$ there is an additional effect that involves the sophisticated buyer: c_e is saved at the cost of a less precise assessment of quality. When present, this positive effect for sophisticated

buyers 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. Table 2 illustrates all these effects. The following proposition summarizes.

PROPOSITION **6.** *(Effects of mandatory labelling on types n and s)*
Let $c_e < \hat{c}_e(p(H))$ and $p(n) < \hat{p}(c_d, p(H))$. Starting from a pooling equilibrium $(\sigma^p, \beta_n^p, \beta_s^p)$, the introduction of mandatory labelling leads to a mandatory equilibrium $(\sigma^m, \beta_n^m, \beta_s^m)$ where the naive buyer is better off. Further, if $p(H|X) < \overline{\mu}(c_e)$ and $p(H|Y) > \underline{\mu}(c_e)$, then the sophisticated buyer neither gains nor loses; otherwise, she is also better off. For both types, the gains (if present) increase monotonically in $p(H|X)$ and decrease monotonically in $p(Y|H)$.

6 Endogenous joint distribution of qualities and traits

The main message that can be extrapolated from the results presented in the previous sections 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 3), then the introduction of mandatory disclosure of the private trait has the consequence of lowering profits for high quality firms (Proposition 4) and, if labels are sufficiently informative on quality, of increasing profits for low quality firms (Proposition 5). Consumers welfare, instead, is not reduced by mandatory disclosure, with naive consumers always gaining and sophisticated consumers gaining when labels are informative enough (Proposition 6). 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 3 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. 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 by sophisticated buyers, in order to guarantee equilibrium existence.[12]

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

[12]The possibility of mixed behavior for buyers does not play any significant role in the model of Section 3, so one can safely neglect it.

different costs. Denote with c_{qt} the fixed cost that must be incurred for producing as type (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.

To allow for mixed behavior among buyers of the same type (e.g., some sophisticated buyers choosing $s1$ and some others choosing $s2$) we normalize the total mass of buyers to 1 and assign to each buyer an index in the interval $[0,1]$, with buyers in $[0,p(s)]$ being sophisticated types and buyers in $(p(s),1]$ being naive types.[13] We also assume that c_e is neither too large nor too small, such that $0 < \underline{\mu}(c_e)$ and $\overline{\mu}(c_e) < 1$, as otherwise sophisticated buyers would always choose $s1$ or always choose $s2$, respectively. For each buyer's type the optimal behavior is again given by Propositions 1 and 2. So, for any belief μ, the optimal behavior of buyer's types can be parsimoniously described by z_μ^* and the total mass of types that acquires q, which we denote with $\lambda \in [0,p(s)]$.

When buyer's beliefs on quality are obtained by means of Bayes' rule and buyers spend optimally, a seller of type (q,t) makes the following profits (per unit-mass of buyers):

$$\pi_{qt} = \begin{cases} \lambda V\left(z_q^*\right) + (1-\lambda) V\left(z_{p(H)}^*\right) - c_{qt} & \text{if t is not disclosed} \\ \lambda V\left(z_q^*\right) + (1-\lambda) V\left(z_{p(H|t)}^*\right) - c_{qt} - c_d & \text{if t is disclosed} \end{cases} \tag{2}$$

In this setup, a profile $(\sigma, \tilde{\beta}_n, \tilde{\beta}_s)$ describes the behavior of all agents' types, with σ a function describing the choice of disclosure for each seller's type (as defined in Section 3), $\tilde{\beta}_s : [0,p(s)] \to \mathcal{B}_s$ a function describing the acquisition and expenditure choices for each sophisticated buyer, where $\mathcal{B}_s$ is the set of all possible functions β_s (as defined in Section 3), and $\tilde{\beta}_n : (p(s),1] \to \mathcal{B}_n$ a function describing the acquisition and expenditure choices for each sophisticated buyer, where $\mathcal{B}_n$ is the set of all possible functions β_n (again, as defined in Section 3). Note that, for given $p(H)$ and $p(H|t)$, $(\sigma, \tilde{\beta}_n, \tilde{\beta}_s)$ describes the behavior of the players of the persuasion game with labelling introduced in Section 3, with two differences: buyers of the same type are allowed to play different strategies and some of the seller's types might not be present.

A *persuasion equilibrium with endogenous seller's types* (PEEST) is a triple $(p(H), (p(H|X), p(H,Y)), (\sigma, \tilde{\beta}_n, \tilde{\beta}_s))$, such that: (i) $(\sigma, \tilde{\beta}_n, \tilde{\beta}_s)$ is a wPBE which survives the D1

[13] This allows to specify a distinct behavior for each $i \in [0,1]$.

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 3, 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 particular, we focus on the PEEST which induces the highest $p(H)$. 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, if sophisticated buyers are many enough, the pooling PEEST with highest $p(H)$ 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 enough sophisticated buyers to acquire q, and hence to sustain the profits of H-types. The following proposition summarizes:

PROPOSITION **7.** *(PEEST with highest $p(H)$ and no voluntary disclosure)*
Let $\lambda^* = \frac{(c_{HX} - c_{LY})}{V(z_H^*) - V(z_L^*)}$. There exists $\check{c}_e > 0$ such that, if $p(n) \leq 1 - \lambda^*$ and $c_e < \check{c}_e$, then the PEEST with highest $p(H)$ is $(\overline{\mu}(c_e), (1,0), (\sigma^p, \tilde{\beta}_n^p, \tilde{\beta}_s^{\lambda^*}))$, where no operating seller's type discloses his own trait, a fraction λ^* of buyers (all sophisticated) acquires q and chooses z_H^* if $q = H$ and z_L^* if $q = L$, while the remaining fraction $1 - \lambda^*$ does not acquire q and chooses $z_{\overline{\mu}(c_e)}^*$. Moreover, if $p(s) < \lambda^*$ then there are no PEEST equilibria with no disclosure and $P(H) > 0$.

The intuition underlying Proposition 7 is simple. No equilibrium with $p(H) > \overline{\mu}(c_e)$ can exist, because for such $p(H)$ no sophisticated 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 $p(H)$, sophisticated buyers are indifferent between acquiring q and not acquiring it; therefore, it can always be found a sufficiently large mass of sophisticated buyers, λ^*, that optimally acquire q and spend z_q^*, keeping H-types on the market.[14] Further,

[14] An equilibrium could exist for $\underline{\mu}(c_e) < p(H) < \overline{\mu}(c_e)$, if there exists a $z_{p(H)}^*$ for which $\pi_{HX} = \pi_{LY}$ when all $p(s)$ acquire q and spend z_q^*. Also, it always exists a PEEST for $p(H) = \underline{\mu}(c_e)$. However, these equilibria are intuitively not stable under reasonable profit-based dynamics, as depicted in Figure 3.

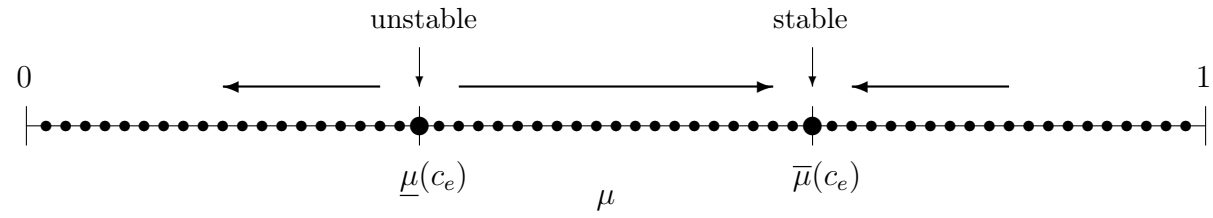

Figure 3: The dynamics of $p(H)$. Consider the case of no disclosure: for $p(H) > \overline{\mu}(c_e)$ or $p(H) < \underline{\mu}(c_e)$ no sophisticated type acquires q, boosting the profits of types L which, under any reasonable profit-based dynamics, leads to a reduction in $p(H)$; for $\underline{\mu}(c_e)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

if $p(n)$ is sufficiently low, then 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 discussion below Proposition 3). 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 maket.

As done in Section 5 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 checking under what respects an improvement is possible (average quality, buyer's utility, production costs). It turns out that mandatory labelling has little effects and, if it has, they might well be undesirable. The following proposition summarizes:

PROPOSITION **8.** *(Effects of mandatory labelling)*
Starting from $(\overline{\mu}(c_e), (1, 0), (\sigma^p, \tilde{\beta}_n^p, \tilde{\beta}_s^{\lambda^*}))$*, the introduction of mandatory labelling can not increase average quality and can not make buyers better off. In addition, if quality does not decrease, then seller's profits can not increase and there are cost-inefficiencies net of disclosure costs.*

The results described in Proposition 8 can be explained as follows. As discussed for the intuition of Proposition 7, in a PEEST the maximum $p(H)$ that can possibly be obtained is equal to $\overline{\mu}(c_e)$. Since in $(\overline{\mu}(c_e), (1, 0), (\sigma^p, \tilde{\beta}_n^p, \tilde{\beta}_s^{\lambda^*}))$ average quality is already at such maximum level, no gain is possibile in this regard. Further, buyers can not gain because their utility is strictly increasing in expected quality and, even if labelling is informative, under mandatory labelling we have that $p(H|t) \leq \overline{\mu}(c_e)$, otherwise type (L, t) would kick

type (H,t) out of the market. Now, consider the case where mandatory labelling does not reduce average quality at all, leaving it at $\overline{\mu}(c_e)$. Such quality is possibly compatible only with certain distributions of seller's types, all of which have the total fraction of H-types equal to $\overline{\mu}(c_e)$. Note that, if type (H,t) is on the market, then also type (L,t) must be on the market, otherwise label t would perfectly signal high quality and no sophisticated buyer would acquire q, with the result that type (L,t) would take over. So, for the presence of type (H,X) to be compatible with a PEEST under mandatory labelling, types (H,X) and (L,X) must earn the same profits, that in turn requires that the mass of sophisticated buyers who acquire q when they see x is smaller than λ^*, since type (L,X) faces greater setup costs than type (L,Y), which in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$ makes the same profits of type (H,X). But if less sophisticated buyers acquire q, then type (H,X) makes smaller profits. Further, for the presence of type (H,Y) to be compatible with a PEEST under mandatory labelling, types (H,Y) and (L,Y) must earn the same profits, that in turn requires that the mass of sophisticated buyers who acquire q when they see y is larger than λ^*, since type (H,Y) faces greater setup costs than type (H,X), which in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$ makes the same profits of type (L,Y). But if more sophisticated buyers acquire q, then type (L,Y) makes smaller profits. A fortiori, types (L,X) and (H,Y) make smaller profits than, respectively, types (L,Y) and (H,X) in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$. Finally, to see why cost-inefficiency must arise, it is sufficient to observe that there is only one distribution of seller's type that is cost-efficient (abstracting from disclosure costs) which is compatible with $p(H)=\overline{\mu}(c_e)$: $p(H|X)=1$ and $p(H|Y)=0$, namely the one that arises in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$. But this distribution is not sustainable as a PEEST under mandatory labelling, because there is perfect correlation between traits and qualities, so that no sophisticated buyer would acquire q, which in turn would wipe H-types out of the market.

7 Conclusions

In this paper we have shown that the presence of sophisticated buyers – with the option to acquire information on product quality at a cost – can lead not only to the failure of market unravelling but also to the backfiring of mandatory disclosure – that a public authority may want to implement as a reaction to the failure of market unravelling. 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.

These results crucially depend on some of the assumptions of the model. One of these is about the information structure: buyers can acquire finer information than what can be certified by labelling. This, of course, is not always the case in real markets. However, in many cases this assumption seems reasonable: actual quality can be deduced from the available information on products (e.g., list of ingredients, details of product design, specialized reviews, data on usage, etc.) but this requires quite an effort on the part of the buyer who has to elaborate properly such information, while some simple and easily understandable label (e.g., no fat/no carb, organic, made in country X, tested by university Y, etc.) can provide useful but less precise information on quality at no cost for the buyer.

Another important assumption on the information structure is the impossibility for buyers to acquire the seller's trait, which can be thought of as due to an excessively high acquisition cost. Such an assumption well fits cases where the knowledge of the trait regards characteristics of the product that are difficult to verify or certify for an individual buyer (e.g., country of origin, adherence to production standard, origin of raw materials, etc.). Instead, when buyers can acquire the trait on their own at a low cost, they can induce (a sort of) market unravelling independently of seller's behavior, leaving no room for mandatory disclosure (if not for shifting costs from buyers to sellers) and making our analysis irrelevant. 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 our main results still hold: market unravelling can fail because sophisticated buyers acquire information on quality and, in such case, mandatory disclosure can backfire.

A further crucial assumption is the presence of disclosure costs. Indeed, if voluntary disclosure were costless for the sellers, then high quality sellers with the best trait would always disclose (to gain from naive buyers) and this would lead to market unravelling. Apart from this, however, the cost of disclosure plays no substantial role. Moreover, in a sense a positive cost of disclosure is the only meaningful assumption if one wants to investigate the desirability of mandatory labelling: if market unravelling takes place, there is evidently no necessity of mandatory disclosure. Incidentally, this also suggests that 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.

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 increasing the number of sophisticated buyers (who are better off with respect to naive ones) and of reducing the cost of information acquisition (which, when market unravelling fails, is paid by all sophisticated buyers). 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' feedbacks.

Acknowledgements

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

References

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.

Bilancini, E. and L. Boncinelli (2014). Persuasion with reference cues and elaboration costs. Technical report.

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.

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

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.

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.

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.

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.

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*.

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.

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.

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

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.

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

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.

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 Proposition 1

It is immediate to observe that $\lim_{\mu\to 0} \mu U(z^*_\mu, H) + (1-\mu)U(z^*_\mu, L) = U(z^*_L, L)$, and $\lim_{\mu\to 0} \mu U(z^*_H, H) + (1-\mu)U(z^*_L, L) = U(z^*_L, L) - c_e$. 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 (s_1, z^*_μ) an optimal for the for the sophisticated buyer. Moreover, if $\mu < \underline{\mu}(c_e)$ then the sophisticated buyer strictly prefers (s_1, z^*_μ).

Analogously, we observe that $\lim_{\mu\to 1} \mu U(z^*_\mu, H) + (1-\mu)U(z^*_\mu, L) = U(z^*_H, H)$, and $\lim_{\mu\to 1} \mu U(z^*_H, H) + (1-\mu)U(z^*_L, L) = U(z^*_H, H) - c_e$. Since the functions involved are continuous, there exists a threshold $\overline{\mu}(c_e)$ such that, if $\mu > \overline{\mu}(c_e)$, then (s_1, z^*_μ) is optimal for the sophisticated buyer. Moreover, if $\mu > \overline{\mu}(c_e)$ then the sophisticated buyer strictly prefers (s_1, z^*_μ).

Proof of Proposition 2

The difference between the maximum expected payoff earned under s_2 and under s_1 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 $c_e \leq \hat{c}_e(\mu)$, then (s_2, z^*_L, z^*_H) is an optimal response for the sophisticated buyer. Moreover, if $c_e < \hat{c}_e(\mu)$, then the sophisticated buyer strictly prefers (s_2, z^*_L, z^*_H).

Proof of Proposition 3

If no trait is observed, then by Bayes rule we have that B must hold a belief equal to $\mu(H) = p(H)$. By Proposition 2 we know that, If $c_e < \hat{c}_e(p(H))$ then (s_2, z^*_L, z^*_H) is the optimal response for the sophisticated buyer. The optimal response for the naive buyer is instead $z^*_{p(H)}$.

Given these responses, sellers of types (H, X) and (H, Y) who do not disclose t obtain a payoff equal to $(1 - p(n))V(z^*_H) + p(n)V(z^*_{p(H)})$. Furthermore, if $p(n) < \hat{p}(c_d, p(H))$, then such payoff is strictly greater than $V(z^*_H) - c_d$ which is the payoff obtained if B holds the belief which is most favorable to S, i.e., $\mu = 1$. Therefore, there is no belief held by B for which a deviation entaling the disclosure of t gives a payoff that is at least as large as $(1 - p(n))V(z^*_H) + p(n)V(z^*_{p(H)})$.

Given B's responses, sellers of types (L,X) and (L,Y) who do not disclose t obtain $(1-p(n))V(z_L^*)+p(n)V(z_{p(H)}^*)$. Let $T\subset[0,1]$ 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 least as large as $(1-p(n))V(z_L^*)+p(n)V(z_{p(H)}^*)$. There are two cases: $T\neq\emptyset$ and $T=\emptyset$. If $T\neq\emptyset$ then, since no belief justifies the disclosure of trait t by types (H,X) and (H,Y), the D1 requires that $\mu(H|x)=\mu(H|y)=0$, and hence by Proposition 1 we know that, when either trait x or trait y is observed, (s_1,z_L^*) is the optimal response for the sophisticated buyer; also, z_L^* is the optimal response for the naive buyer. As a consequence, types (L,X) and type (L,Y) can gain by disclosing t at most a payoff equal to $V(z_L^*)-c_d$, which makes deviating not profitable. If, instead, $T=\emptyset$ then, by the same token, the D1 criterion allows that $\mu(H|x)=\mu(H|y)=p(H)$, and hence the optimal response for the buyer who observes x or y is the same as when nothing is observed. Since $c_d>0$, again neither type (L,X) nor type (L,Y) can gain by disclosing t.

Proof of Proposition 4

Since $c_e<\hat{c}_e(p(H))$ and $p(n)<\hat{p}(c_d,p(H))$, by Proposition 3 we have that $(\sigma^p,\beta_n^p,\beta_s^p)$ exists where the sophisticated buyer acquires q; then, $(\sigma^m,\beta_n^m,\beta_s^m)$ is well defined.

Since in the pooling equilibrium $(\sigma^p,\beta_n^p,\beta_s^p)$ the sellers of types (H,t), $t=X,Y$, are recognized as high quality by the sophisticated buyer, these seller's types obtain a payoff equal to $(1-p(n))V(z_H^*)+p(n)V(z_{P(H)}^*)$. In the mandatory labelling equilibrium $(\sigma^m,\beta_n^m,\beta_s^m)$ the sellers of types (H,t), $t=X,Y$, obtain a payoff equal that is either equal to $(1-p(n))V(z_H^*)+p(n)V(z_{P(H|t)}^*)-c_d$, if the sophisticated buyer still acquires q, or equal to $V(z_{P(H|t)}^*)-c_d$, if the sophisticated buyer does not acquire q. The following inequalities show that both payoffs under $(\sigma^m,\beta_n^m,\beta_s^m)$ are strictly lower than the payoff under $(\sigma^p,\beta_n^p,\beta_s^p)$:

$$\begin{aligned}
(1-p(n))V(z_H^*)+p(n)V(z_{P(H)}^*)-c_d<\\
<(1-p(n))V(z_H^*)+p(n)V(z_{P(H)}^*)-p(n)[V(z_H^*)-V(z_{p(H)}^*)]=\\
=(1-2p(n))V(z_H^*)+2p(n)V(z_{p(H)}^*)<(1-p(n))V(z_H^*)+p(n)V(z_{P(H)}^*) \qquad (3)\\
V(z_{P(H|t)}^*)-c_d<V(z_{P(H|t)}^*)-p(n)[V(z_H^*)-V(z_{p(H)}^*)]=\\
=V(z_{P(H|t)}^*)-p(n))V(z_H^*)+p(n)V(z_{p(H)}^*)<(1-p(n))V(z_H^*)+p(n)V(z_{P(H)}^*) \qquad (4)
\end{aligned}$$

where the first inequality in both (3) and (4) holds since $p(n)<\hat{p}(c_d,p(H))$ implies that $c_d>p(n)[V(z_H^*)-V(z_{p(H)}^*)]$, the second inequality in (3) holds because $z_H^*>z_{P(H)}^*$, and the second inequality in (4) holds because $z_H^*>z_{P(H|t)}^*$, $t=X,Y$.

Proof of Proposition 5

Since $c_e < \hat{c}_e(p(H))$ and $p(n) < \hat{p}(c_d, p(H))$, by Proposition 3 we have that $(\sigma^p, \beta_n^p, \beta_s^p)$ exists where the sophisticated buyer acquires q; then, $(\sigma^m, \beta_n^m, \beta_s^m)$ is well defined.

Consider the seller of type (L, X). Since in the pooling equilibrium $(\sigma^p, \beta_n^p, \beta_s^p)$ the type (L, X) is recognized as low quality by the sophisticated buyer, he obtains a payoff equal to $(1 - p(n))V(z_L^*) + p(n)V(z_{P(H)}^*)$. In the mandatory labelling equilibrium $(\sigma^m, \beta_n^m, \beta_s^m)$, if $p(H|X) > \overline{\mu}(c_e)$, then the sophisticated 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 $V(z_{P(H|X)}^*) - c_d > (1 - p(n))V(z_L^*) + p(n)V(z_{P(H)}^*)$, which gives the result.

Consider the seller of type (L, Y). The result follows from the same argument.

Proof of Proposition 6

Since $c_e < \hat{c}_e(p(H))$ and $p(n) < \hat{p}(c_d, p(H))$, by Proposition 3 we have that $(\sigma^p, \beta_n^p, \beta_s^p)$ exists where the sophisticated buyer acquires q; then, $(\sigma^m, \beta_n^m, \beta_s^m)$ is well defined.

Consider the buyer of type n. In $(\sigma^p, \beta_n^p, \beta_s^p)$ her expected payoff is equal to:

$$p(H)U(z_{p(H)}^*, H) + (1 - p(H))U(z_{p(H)}^*, L) \tag{5}$$

We observe that this payoff does not depend on $p(H|t)$, $t = X, Y$. In $(\sigma^m, \beta_n^m, \beta_s^m)$ her expected payoff is equal to:

$$\begin{aligned} &p(X)[p(H|X)U(z_{p(H|X)}^*, H) + (1 - p(H|X))U(z_{p(H|X)}^*, L)] + \\ &\quad p(Y)[p(H|Y)U(z_{p(H|Y)}^*, H) + (1 - p(H|Y))U(z_{p(H|Y)}^*, L)] \end{aligned} \tag{6}$$

We note that as $p(H|X)$ and $p(H|Y)$ tend to $p(H)$, (6) tends to (5). Moreover, taking into account the optimality of $z_{p(H|X)}^*$ and $z_{p(H|Y)}^*$ and the fact that $\mathrm{d}p(H|Y) = -\mathrm{d}p(H|X)p(X)/p(y)$, the derivative of (6) with respect to $p(H|X)$ is equal to:

$$\begin{aligned} &p(X)[U(z_{p(H|X)}^*, H) - U(z_{p(H|X)}^*, L)] - p(X)\left[U(z_{p(H|Y)}^*, H) - U(z_{p(H|Y)}^*, L)\right] = \\ &= p(X)[U(z_{p(H|X)}^*, H) - U(z_{p(H|Y)}^*, H)] + p(X)[U(z_{p(H|X)}^*, L) - U(z_{p(H|Y)}^*, L)] \end{aligned} \tag{7}$$

We observe that both terms of the right hand side of (7) are positive, since $z_{mu}^* > 0$ for all $\mu \in [0, 1]$ and $\partial U(z, H)/\partial z > \partial U(z, L)/\partial z$. Hence, (6) is strictly greater than (5) if $p(H|X) > p(H) > p(H|Y)$ and it increases in monotonically in $p(H|X)$ and decreases monotonically in $p(H|Y)$.

Consider the buyer of type s. In $(\sigma^p, \beta_n^p, \beta_s^p)$ her expected payoff is equal to:

$$p(H)U(z_H^*, H) + (1 - p(H))U(z_L^*, L) - c_e \tag{8}$$

We observe that this payoff does not depend on $p(H|t)$, $t = X, Y$. In $(\sigma^m, \beta_n^m, \beta_s^m)$ her expected payoff is also equal to (8) if $p(H|X) < \overline{\mu}_{c_e}$ and $p(H|Y) > \overline{\mu}_{c_e}$, because s_2 is still an optimal response. Hence, mandatory labelling provides no benefit to her. Otherwise, s_1 is an optimal response at least when one of the two traits is observed. If it is an optimal response for both trait, then her payoff is equal to or greater than (6), and the argument described above for type n applies to type s too. If s_1 is an optimal response only when one of the two traits is observed, then her payoff is a convex combination of (8) and (6), and again we can apply the argument described above for type n.

Proof of Proposition 7

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 Proposition 1 and Bayes rule, all sophisticated buyers must optimally respond with $\beta_s^p(t) = (s_1, z^*_{p(H|t)})$. So, if t is disclosed, then type (L, t) makes greater profits than type (H, t) which, by condition (ii), implies than $p(H|t) = 0$. 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 Proposition 1 and Bayes rule, all sophisticated buyers must optimally respond with $\beta_s^p(0) = (s_1, z^*_{p(H)})$. This in turn implies that L-types always make greater profits than H-types which, by condition (ii), implies than $p(H) = 0$.

Now, we show that $(\sigma^p, \tilde{\beta}_n^p, \tilde{\beta}_s^{\lambda^*})$ 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 buyers given σ^p and $p(H) = \overline{\mu}(c_e)$, and from this we show that the set of beliefs T_H that justifies a deviation by a H-type is either empty or strictly contained in the set T_L that justifies a deviation by a L-type.

From σ^p follows that no trait is observed. By Bayes rule, a buyer's belief conditional to not observing any label must be equal to $\overline{\mu}(c_e)$. Therefore, by means Proposition 1 and 2 it is straightforward to conclude that a sophisticated buyer is indifferent between (s_2, z_L^*, z_H^*) and $(s_1, z^*_{\overline{\mu}(c_e)})$. Hence, any $\tilde{\beta}_s$ such that $\tilde{\beta}_s(i) = \beta_n^p$ or $\tilde{\beta}_s(i) = \beta_s^p$ induces optimal responses. If $p(s) \geq \lambda^*$ then this holds in particular for any $\tilde{\beta}_s^{\lambda^*}$ where a mass λ^* of sophisticated buyers plays $\beta_s^p(0) = (s_2, z_L^*, z_H^*)$ while a mass $p(s) - \lambda^*$ of them plays $\beta_s^p(0) = (s_1, z^*_{\overline{\mu}(c_e)})$. The optimal response for all naive buyers is $\beta_n^p(0) = (s_1, z^*_{\overline{\mu}(c_e)})$.

Given these responses, 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 $(1 - p(n)) V(z_H^*) + p(n) V(z_\mu^*) - c_d$ if $\mu \in [\underline{\mu}(c_e), \overline{\mu}(c_e)]$, because

all sophisticated buyers can optimally choose (s_2, z_L^*, z_H^*) (from Proposition 2), and equal to $V(z_\mu^*) - c_d$ if $\mu \in [0, \underline{\mu}(c_e)) \cup (\overline{\mu}(c_e), 1]$, because all sophisticated buyers optimally choose (s_1, z_μ^*) (by Proposition 1). Consider the case of $\mu \in (\underline{\mu}(c_e), \overline{\mu}(c_e))$. Then, not disclosing is strictly better than disclosing if:

$$\lambda^* V(z_H^*) + (1-\lambda^*) V(z^*_{\overline{\mu}(c_e)}) > (1-p(n)) V(z_H^*) + p(n) V(z_\mu^*) - c_d \tag{9}$$

which is implied by:

$$\begin{aligned} \lambda^* V(z_H^*) + (1-\lambda^*) V(z^*_{\overline{\mu}(c_e)}) &> (1-p(n)) V(z_H^*) + p(n) V(z^*_{\overline{\mu}(c_e)}) - c_d \quad \Leftrightarrow \\ V(z_H^*) - V(z^*_{\overline{\mu}(c_e)}) &< \frac{c_d}{1-p(n)-\lambda^*} \end{aligned} \tag{10}$$

since $p(s) = 1 - p(n) > \lambda^*$ and $V(z^*_{\overline{\mu}(c_e)}) > V(z_\mu^*)$ for all $\mu \in [(\underline{\mu}(c_e), \overline{\mu}(c_e)]$. Further, since V is strictly increasing in z, z_μ^* is strictly increasing in μ and $\overline{\mu}_{c_e}$ is strictly increasing in c_e, it follows that there always exists $\check{c}_e > 0$ such that the sides of (10) are equal. 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 $\mu \in [0, \underline{\mu}(c_e)) \cup (\overline{\mu}(c_e), 1]$. For such beliefs, not disclosing is strictly better than disclosing if:

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

Since $V(z_H^*) \geq [\lambda^* V(z_H^*) + (1-\lambda^*) V(z^*_{\overline{\mu}(c_e)})] \geq V(z^*_{\overline{\mu}(c_e)})$ and $V(z_\mu^*)$ is strictly increasing in μ and tends to $V(z^*_{\overline{\mu}(c_e)})$ for μ which tends to $\overline{\mu}(c_e)$ and to $V(z_H^*)$ for μ which tends to 1, it follows that (11) is satisfied if and only if $\mu \in T_H = (\overline{\mu}(c_e) + \epsilon, 1]$, where $\overline{\mu}(c_e) + \epsilon$ equates the two sides of (11).

Given B's responses, 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 if $\mu \in T_H$ then $\mu \in T_L$. Moreover, since $V(z_L^*) \leq [\lambda^* V(z_L^*) + (1-\lambda^*) V(z^*_{\overline{\mu}(c_e)})] \leq V(z^*_{\overline{\mu}(c_e)})$ and $V(z_\mu^*)$ is strictly increasing in μ and tends to $V(z^*_{\overline{\mu}(c_e)})$ for μ which tends to $\overline{\mu}(c_e)$ and to $V(z_L^*)$ for μ which tends to 0, we have that if $T_L \neq \emptyset$ then there exists $\epsilon' > 0$ such that, for $\mu \in (\overline{\mu}(c_e) + \epsilon', \overline{\mu}(c_e) + \epsilon)$, the following inequalities hold:

$$\lambda^* V(z_L^*) + (1-\lambda^*) V(z^*_{\overline{\mu}(c_e)}) < V(z_\mu^*) - c_d < \lambda^* V(z_H^*) + (1-\lambda^*) V(z^*_{\overline{\mu}(c_e)}) \tag{12}$$

which in turn implies that T_H is strictly contained in T_L and therefore the D1 criterion requires that $\mu(H|x) = \mu(H|y) = 0$, making the disclosure of t unprofitable for all seller's types. If, instead, $T_L = \emptyset$ then, by the same token, also $T_H = \emptyset$, and the D1 criterion allows that $\mu(H|t) = p(H)$, and hence the optimal response for the buyer who observes x or y is the same as when nothing is observed. Since $c_d > 0$, again no seller's type can gain by disclosing t.

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

$$\pi_{HX} = \lambda^* V(z_H^*) + (1-\lambda^*)V(z^*_{\overline{\mu}(c_e))}) - c_{HX} = \lambda^* V(z_L^*) + (1-\lambda^*)V(z^*_{\overline{\mu}(c_e)}) - c_{LY} = \pi_{LY} \;\Leftrightarrow$$
$$\Leftrightarrow\; \lambda^* V(z_H^*) - c_{HX} = \lambda^* V(z_L^*) - c_{LY} \tag{13}$$

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

Finally, we show that $p(s) \geq \lambda^*$ is a necessary condition for the existence of PEEST equilibria with no disclosure and $p(H) > 0$. Suppose that $p(s) < \lambda^*$. Then, there is no way to satisfy (13), as it is independent of $\overline{\mu}(c_e)$. In particular, type (L, Y) always makes strictly more profits than any other seller's types, ruling out the possibility of a PEEST with no disclosure and $P(H) > 0$.

Proof of Proposition 8

Firstly, 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 obtain. This is evident if one particularizes the argument provided in the first paragraph of the proof of Proposition 7 to the case of $\sigma = \sigma^m$, i.e., abstracting from cases where buyers do not observe any label.

Conditional on observing t, the utility of naive buyers strictly and positively depends on $p(H|t)$, $t \in \{X, Y\}$, while under no disclosure it strictly and positively depends on $p(H)$ only. Moreover, the two utilities coincide when $p(H) = p(H|t)$. In $(\overline{\mu}(c_e), (1,0), (\sigma^p, \tilde{\beta}_n^p, \tilde{\beta}_s^{\lambda^*}))$ there is no disclosure and $p(H) = \overline{\mu}(c_e)$ while, by the argument above, under mandatory disclosure $p(H|t) \leq p(H)$, so that the utility of naive buyers can not increase. To see that the same result holds for the utility of sophisticated buyers, it suffices to note that when sophisticated buyers are best responding their utility conditional to observing t is strictly increasing in $p(H|t)$, $t \in \{X, Y\}$, while under no disclosure it strictly and positively depends on $p(H)$ only.

Finally, suppose that $(\overline{\mu}(c_e), (p(H|X), p(H|Y)), (\sigma^m, \tilde{\beta}_n^m, \tilde{\beta}_s^m))$ is a PEEST under mandatory labelling. Since $p(H) = \overline{\mu}(c_e)$, then only the following values are possible for $(p(H|X), p(H|Y))$:

$(1,0)$, $(1,\emptyset)$, $(0,1)$, $(\emptyset,1)$, $(\overline{\mu}(c_e),\overline{\mu}(c_e))$, $(\overline{\mu}(c_e),\emptyset)$, or $(\emptyset,\overline{\mu}(c_e))$. However, $(p(H|X),p(H|Y)) = (1,0)$ is impossible, because by Proposition (1) sophisticated buyers would never acquire q, which is necessary for $p(H|X)=1$. A similar argument rules out $(p(H|X),p(H|Y))=(0,1)$. Let us consider seller's profits. If $p(H|X)=\overline{\mu}(c_e)$, then type (H,X) can not make greater profits than what made in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$, because the fraction of buyers acquiring q after observing label x will be strictly lower than λ^*, since in this case $\pi_{HX}=\pi_{LX}$ while in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$ it must hold $\pi_{HX}=\pi_{LY}$ with $c_{LX}>c_{LY}$, and in addition expected quality for buyers not acquiring q also remains the same. If $p(H|Y)=\overline{\mu}(c_e)$, then type (L,Y) can not make greater profits than what made in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$, because the fraction of buyers acquiring q after observing label y will be strictly greater than λ^*, since in this case $\pi_{HY}=\pi_{LY}$ while in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$ it must hold $\pi_{HX}=\pi_{LY}$ with $c_{HY}>c_{HX}$, and in addition expected quality for buyers not acquiring q also remains the same. From this argument also follows that types (H,Y) and (L,X) can not earn more profit than what earned, respectively, by (H,X) and (L,Y) in in $(\overline{\mu}(c_e),(1,0),(\sigma^p,\tilde{\beta}_n^p,\tilde{\beta}_s^{\lambda^*}))$. A similar argument shows that the same holds for types (L,t). If $p(H|t)=0$, then only type (L,t) operates and, by Proposition 1, sophisticated buyers never acquire q upon observing t, with the result that type (L,t) makes the minimum possible profits. If $p(H|t)=$, then no t-type operates and, hence, profits are 0. Let us consider to cost-efficiency. Since $c_{HX}<c_{HY}$ and $c_{LX}>c_{LY}$, cost-efficiency requires that, if both quality H and quality L are produced, we have $(p(H|X),p(H|Y))=(1,0)$, which is impossible by argument above.