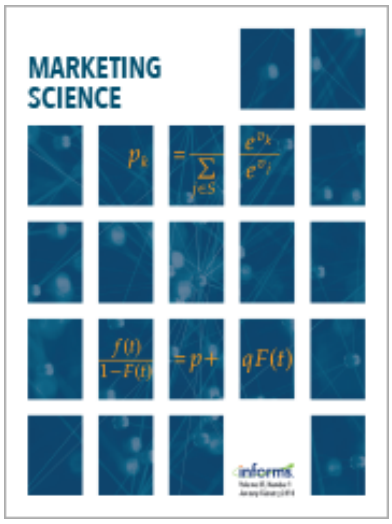



Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Product Line Design Under Preference Uncertainty Using Aggregate Consumer Data

Zibin Xu, Anthony Dukes

To cite this article:
Zibin Xu, Anthony Dukes (2019) Product Line Design Under Preference Uncertainty Using Aggregate Consumer Data. Marketing Science 38(4):669-689. https://doi.org/10.1287/mksc.2019.1160

http://pubsonline.informs.org/journal/mksc/

MARKETING SCIENCE
Vol. 38, No. 4, July–August 2019, pp. 669–689
ISSN 0732-2399 (print), ISSN 1526-548X (online)

# Product Line Design Under Preference Uncertainty Using Aggregate Consumer Data

**Zibin Xu,[a] Anthony Dukes[b]**

[a] Antai College of Economics and Management, Shanghai Jiao Tong University, 200030 Shanghai, China; [b] Marshall School of Business, University of Southern California, Los Angeles, California 90089

**Contact:** zibin.xu@gmail.com, http://orcid.org/0000-0002-0001-2365 (ZX); dukes@marshall.usc.edu, http://orcid.org/0000-0003-2699-8019 (AD)

**Received:** February 10, 2018
**Revised:** October 1, 2018; November 24, 2018
**Accepted:** December 13, 2018
**Published Online in Articles in Advance:** July 12, 2019

**https://doi.org/10.1287/mksc.2019.1160**



**Abstract.** This research studies the product line design problem when consumers are subject to perceptual errors in assessing their intrinsic preferences. If perceptual errors are driven by common variables, then a firm may use aggregate consumer data (e.g., conjoint studies or anonymous usage data) to deduce the errors and infer the consumer preferences. In this way, we develop microfoundations necessary to show when and how the firm can understand consumer preferences better than consumers themselves, a situation we call *superior knowledge*. But is superior knowledge ever unprofitable? How should the firm with superior knowledge design its product line? Do consumers receive more-relevant products or simply have more surplus extracted? Can data collection help consumers make better choices? Our results suggest that consumers' rational suspicions may prevent the firm from exploiting its superior knowledge. In addition, the burden of signaling may force the firm to offer efficient quality for its products. Therefore, allowing the firm to collect aggregate consumer data may be strictly Pareto improving.

**History:** Yuxin Chen served as the senior editor and Juanjuan Zhang served as associate editor for this article.
**Funding:** This work was supported by National Natural Science Foundation of China [Grant 71802131].
**Supplemental Material:** The online appendix is available at https://doi.org/10.1287/mksc.2019.1160.

**Keywords:** consumer data collection • product line design • superior knowledge • uninformed preference • perceptual error • signaling model

## 1. Introduction

This paper studies product line design when the firm knows about consumers' preferences better than some do about themselves. These situations are possible for new products with innovative features. Consider the case of the Aeron chair developed by the office furniture manufacturer Herman Miller in 1992. Surveys and marketing research instructed designers on what features to build into the chair, yet consumers' initial impressions were negative. But by 2010, the Aeron chair was the best-selling office chair of all time. That is, despite consumers' pessimistic perceptions, the Aeron chair was marketed with the confidence that consumers will eventually appreciate the product, even before consumers do.

For new products, a consumer's initial perception may be an imperfect assessment of her intrinsic, ex post, preference.[1] The market often has a hard time appreciating innovative features, because consumers' perceptions may be susceptive to their noisy prior experiences with stale product categories. Initial assessments of the Aeron chair, for example, may have been imperfect because of the influences from the older models of office chairs available on the market. When consumers' *perceptual noise* is driven by external market factors, a sample of consumer data can uncover the systematic noise to reveal the market's latent preferences. Even if the data are anonymous, viewing the data in aggregate permits marketing researchers to deduce perceptual noises by studying the pattern of consumer reactions to the market factors and to ultimately understand consumers' intrinsic preferences. In such a case, we say the firm has *superior knowledge* of consumer preferences.

As data technologies improve, the situation of superior knowledge becomes increasingly relevant because firms have ever more access to individual data. Such situations are evident in practice. For instance, one headline from *Variety* states that "Netflix knows its subscribers better than they do" (Roettgers 2018), and author-economist Seth Stephens-Davidowitz (2017) quotes a former data scientist at a major digital content company: "The algorithms [of Big Data] know you better than you know yourself" (p. 157). Although consumer protection advocates raise worries that data collection will subject consumers to exploitation,[2] marketers argue that consumer data help them offer better-suited products and services. In this research, we contribute to this debate by asking the following questions: Is data collection ever unprofitable? How

should a firm with superior knowledge design its product line to consumers with perceptual errors? Does the firm offer better-suited products? Or does the firm exploit consumers' perceptual errors? Can uninformed consumers learn their intrinsic values and make better choices by observing the firm's product menu?

To address the above questions, we develop the microfoundations necessary to show how and when the firm may acquire superior knowledge. We are particularly interested in a firm's inferences from anonymous consumer data, which are collected through either traditional market research methods (e.g., subjective data from surveys or conjoint studies) or more sophisticated data collection techniques (e.g., online usage data tracking). The data, even though collected anonymously, can help marketers estimate the perceptual errors in the market. In addition, we discuss how superior knowledge may confound the product line design problem. As is well known, when consumers are perfectly informed about their preferences, the firm can ensure proper self-selection through the downward distortion of quality at the bottom of the product line (Mussa and Rosen 1978, Moorthy 1984). But if some consumers are imperfectly informed of their preferences, how can the firm convince them to sort themselves across the menu of product line options? Overcoming this obstacle is the fundamental challenge of this research.

To see how data aggregation may create superior knowledge, consider again the office chair example above. Suppose there are two segments of consumers that differ by the relative strength of their intrinsic preferences for the Aeron chair ($H$-types and $L$-types). A sample of consumers is surveyed for their perceptions of the product prototype. Each consumer's perception is generated jointly by her intrinsic preference and contextual influences, such as office furniture available in the past. Consumers with a single observation may not necessarily reveal whether their low perception is generated by the low intrinsic preference or by the pessimistic market state. However, by viewing the aggregate sample, the firm can obtain superior knowledge on the consumers' intrinsic preferences. For instance, consider the following numerical example. Suppose that the $H$-types' intrinsic value for the product is $\alpha_H = 3$, and the $L$-types' intrinsic value is $\alpha_L = 1$. Let consumer perception be $\theta_{it} = \alpha_i + \beta_t$, where $\beta_t$ refers to the transitory market state, a random variable of either optimism or pessimism (i.e., $\beta_O = 1$ or $\beta_P = -1$, respectively). Suppose years of uninspiring and routine office furniture indicate a pessimistic *market state* (i.e., $\beta_O = -1$), which has negatively influenced consumers' perceptions. Then a consumer with the "neutral" perception of $\theta = 2$ is unsure whether she is an $H$-type or an $L$-type, because both $\theta_{LO}$ and $\theta_{HP}$ equal 2. But the firm observing the aggregate the data $\{2, 0\}$ learns the pessimistic state and thus infers that the neutral consumers are $H$-types, who are underestimating their intrinsic value.

The premise of ambiguous intrinsic preferences is consistent with the behavioral science literature on *inherent preferences* (Simonson 2008), which views consumers' behaviors as driven by consistent patterns that characterize their unique intrinsic preferences that may not be perfectly recallable. Although we maintain the inherent preferences interpretation, it is important to recognize that one can, alternatively, interpret the aggregation mechanism by the *constructed preferences* view (Bettman et al. 1998, 2008), which argues that consumers develop new preferences over time. In this way, consumer perceptions reflect their true preferences—but the preferences are temporary and susceptible to transitory environmental factors (market state). Under this alternative view, our mechanism suggests that the firm may better predict the consumers' future expected constructed preferences by deducing the environmental influences in their current preferences.

Knowledge of the market state is valuable for the firm, because it reveals what consumers believe about their true preferences. But any rational consumer who is uninformed of her type will be suspicious of overpaying, because the firm has an incentive to mislead the uninformed $L$-types to overbuy the better-quality product with a higher price. The challenge for the firm with superior knowledge is to design the product line while managing these suspicions. Specifically, uninformed consumers will try to update their beliefs by observing the product menu. In this way, decisions about the product menu constitute a signaling game between the firm with private information and the uninformed consumers. Our model identifies a unique separating equilibrium in which consumer suspicions prevent the firm from tricking them to overpay.

Our analysis indicates that the constraint imposed by consumer suspicions may supplant the classical cannibalization constraint, inducing the firm to design a more efficient product line than in the standard model (Mussa and Rosen 1978, Moorthy 1984). In contrast to the standard model, the firm with superior knowledge must design the product line to induce a priori uninformed $H$-types to buy the high-end product. To convince them that they are not being tricked into overbuying a mismatched product, the firm signals its private information, the unobserved market state, by setting the price of the high-end product lower than in the standard model. The reduction in price of the high-end product relaxes the cannibalization constraint, because the $H$-type consumers would have less incentive to buy the cheaper alternative. Thus, the quality of the lower-end

product can be raised to the efficient level. The additional surplus created by the efficiency restoration becomes a source of profits for the firm and compensates for the reduction in the high-end product's price. Therefore, data collection may be a strict Pareto improvement.

The return of surplus to the *H*-type consumers represents the firm's signaling cost when communicating private information about the market state (e.g., Milgrom and Weber 1985). The burden of signaling can be so significant that, under some conditions, the firm would prefer not to collect data. In this case, the firm would design a product menu based on the prior distribution of consumers' initial perceptions. The product menu without data collection is inefficient, because the firm will further downward distort the low-end product quality to prevent cannibalization in the event that *H*-types are underestimating their value.

Previous research on data collection suggests that consumers are typically exploited when a firm has access to information on their willingness to pay. Therefore, without further inducements, consumers are often worse off when data collection occurs, especially those in the high-end market segment[3] (e.g., Villas-Boas 2004, Acquisti and Varian 2005). However, from the above discussions, we identify three distinctive beneficial roles of data collection for consumers. First, data collection may help uninformed consumers learn their intrinsic values and make better choices. Specifically, *L*-types can avoid overbuys and *H*-types can avoid underbuys. Second, data collection enables a product menu that better fits with the needs of the market. Without data collection, the firm cannot be accurate about the distribution of consumers' perceptions and, as a result, design poorly suited products with little sales. Third, data collection may force the firm to amend the product menu and then, as a result of the signaling cost, cut the price of the high-end product, which leaves additional surplus to uninformed *H*-types. Note that the root cause of these three benefits is that data collection may create superior knowledge and therefore invoke consumers' suspicions, which then forces the firm to convince the uninformed consumers of their type by signaling.

Our findings also contrast with classic product line design literature with respect to the welfare implications of second-degree price discrimination. Mussa and Rosen (1978) suggests that second-degree price discrimination by a monopolist is always profitable. Stokey (1979) and Anderson and Dana (2009) show that a monopolist may forgo product versioning only if there exist constraints on technology capacities or product qualities. By contrast, our research suggests that even without these constraints, the monopolist may find it unprofitable to enhance its ability to second-degree price discriminate. In addition, although the existing models suggest that endowing the firm with the ability to price discriminate may lead to a Pareto improvement (Varian 1985, Anderson and Dana 2009), this result occurs only under the condition that price discrimination enables the firm to serve more consumers. Furthermore, the Pareto improvement is always weak, leaving at least some consumers with no additional surplus. By contrast, our research shows that second-degree price discrimination with superior knowledge can be a strict Pareto outcome, both ex ante and ex post, even without the output expansion.

In addition, our research extends the literature on consumer preference discovery (Wernerfelt 1995, Kamenica 2008, Guo and Zhang 2012) by demonstrating that uninformed consumers may infer their type from product lines. Specifically, Guo and Zhang (2012) considers consumer deliberation as independent efforts with exogenous costs. By contrast, our research examines deliberation as an inference process that takes part in the strategic interaction with the firm, allowing the possibility that the firm may manipulate consumers' beliefs. Furthermore, our model requires less restrictions on consumers' prior knowledge. For example, Wernerfelt (1995) and Kamenica (2008) assume that consumers are uninformed of the absolute value of their willingness to pay, but they need to know ex ante their relative ranking in the market. By contrast, in our model an uninformed consumer need not know her ranking in the market. Relaxing this restriction enables us to develop the microfoundations for superior knowledge.

Finally, our research offers two different insights to the recent discussions about consumers' private information (Varian 2002, Taylor 2004, Calzolari and Pavan 2005, Fudenberg and Villas-Boas 2006). First, the literature focuses on price exploitation rather than product design, suggesting that data collection strictly reduces consumers' average welfare by depriving consumers any information rent. This result holds only when consumers know their preferences ex ante, which assumes that the only role of data collection is to identify consumers for price exploitation. But when superior knowledge is possible, data collection can also help the firm design better-suited products. Therefore, the trade-off between product designs and price discrimination requires formal examination. Second, unlike the literature that characterizes data collection as a one-way information transmission from consumers to the firm, our model considers the situations in which data collection may become a two-way information exchange (Samuelson 2000, Solove 2007): It starts with the firm learning from consumers by observing their data, which, by data aggregation, creates

superior knowledge that exceeds uninformed consumers' self-awareness. Then, uninformed consumers learn from the firm through its product line design when it credibly signals its superior knowledge. The dual-directional role of data collection in helping uninformed consumers learn their preference is not seen in earlier literature.

In the following section, we lay out the basics of a general model. In Section 3, we specify the conditions needed to generate the firm's superior knowledge and discuss the role of data aggregation in market research. In Section 4, we provide a stylized numerical example to illustrate the central mechanism of our model. Section 5 is a comprehensive description of the model's equilibrium properties, followed by a presentation of our main results in Section 6. Section 7 concludes. All proofs and formal results are provided in the appendix.[4]

## 2. The General Model

A firm designs a menu of vertically differentiated products to serve two types of consumers (normalized to unitary measure in total), indexed by $i \in \{H, L\}$. A fraction $\lambda \in (0, 1)$ of consumers are $H$-types, and the others are $L$-types. The products are differentiated in quality $q_k$ and corresponding price $p_k$, where $k$ refers to the quality rank of products. We will denote the order of quality as $q_1 > q_2 > q_3 > \cdots$, without loss of generality. For each unit of product $k$, the firm incurs a production cost of $(\frac{1}{2}q_k^2)$. All consumers observe the product line (the price and quality of all available products) without ambiguity and may choose one or no product to purchase. A consumer $i$ has an intrinsic marginal utility for product quality $\alpha_i$, which we refer to as her intrinsic *consumer value*. Therefore, consumer $i$'s utility from product $k$ is $\alpha_i q_k - p_k$, where $\alpha_L < \alpha_H$. The starting point for this model is found in Mussa and Rosen (1978) and Moorthy (1984), in which consumers are perfectly informed of their value.

We depart from the classic model by assuming that consumer $i$ may have imperfect knowledge of her type. She observes a private signal, which is a noisy indication of her value. Specifically, consumer $i$ receives the signal $\theta_i(m_i)$, which is a differentiable function of the noise parameter, $m_i \sim U(0, 1)$. We refer to $m_i$ as consumer $i$'s *perceptual errors*. One can interpret this $\theta_i$ as consumer $i$'s initial assessment about a new product after a prepurchase trial. Perceptual errors $m_i$ confound the consumer's estimation of her type $\alpha_i$. A consumer's inability to initially determine the strength of her preference for a new product may be caused by common market factors. One such factor we have in mind is status quo bias (Samuelson and Zeckhauser 1988). For example, when consumers are accustomed to older office chair models, they may not immediately appreciate the innovative aspects of the newer models. Alternatively, overestimation could result when the consumer is subject to market hype associated with, for example, informational cascading (Anderson and Holt 1997). The fact that consumers are subject to these environmental factors does not presume that they are irrational. In fact, any consumer is eventually aware that her intrinsic preferences may be initially obscured.

We suppose that the firm can choose to observe the consumers' signals $\{\theta_L, \theta_H\}$ before specifying the qualities and prices. We interpret the firm's observation of these signals as data from marketing research conducted at the individual consumer level without personal identification information. It may be helpful to imagine, for example, the firm conducting a survey (e.g., conjoint) of consumers' reactions[5] to a prepurchase trial or product prototype. We refer to $\theta_i$ as *individual consumer data* and the collected data set $\{\theta_L, \theta_H\}$ as *aggregate consumer data*. Under specific conditions discussed in the next section, aggregate consumer data reveal information about individual consumers that extends beyond what consumers know about themselves individually—a situation of *superior knowledge* for the firm.

The game unfolds over three periods. In period 0, nature randomly draws $m_i$, and each consumer $i$ observes her own data $\theta_i$. In period 1, the firm chooses whether to collect consumer data $\{\theta_L, \theta_H\}$. Consumers observe whether the firm collects data. In period 2, the firm designs its product line $\{q_k, p_k\}_{k=1,2,3...}$; then each consumer observes the product line and makes a purchase decision.

## 3. Superior Knowledge and the Role of Data Aggregation

Because both types of consumers may observe the same $\theta$ under some market state, it is possible that a consumer may not learn her type from observing her signal. We call such a consumer *uninformed*.

**Definition 1.** Consumer $i \in \{L, H\}$ is *uninformed* if there exists $m, m' \in (0, 1), m \neq m'$, such that $\theta_i(m) = \theta_j(m')$ for $i \neq j$. Otherwise, consumer $i$ is *informed*.

As formally expressed in this definition, the uninformed consumer realizes that the same signal can be observed by either type depending on the market state, and therefore, she cannot assess her type from her signal. Definition 1 also presumes the possibility that some consumers may be initially informed by their own signals. For instance, suppose that $m$ is such that $\theta_L(m) < \theta_H(m')$ for all $m' \in (0, 1)$. The perception of the product is so low that she deduces that not even under the most pessimistic market state would $H$-types have such a low perception. In this case, a consumer is informed that she is $L$-type.

### 3.1. From Data Aggregation to Superior Knowledge

Because aggregate data enable the firm to observe more data than any consumer does, the firm may deduce the consumers' perceptual errors and thus obtain informational advantage over the consumers who are uninformed of their type. In this section, we identify the conditions in which the firm may obtain an informational advantage by aggregating consumer data. We start by specifying what we mean by an informational advantage.

**Definition 2.** The firm has *superior knowledge* over uninformed consumer $i$, if it deduces her true type by observing the aggregate data $\{\theta_L, \theta_H\}$.

The role of superior knowledge in our results is seen most starkly by starting with the case without it. For instance, consider a case when data aggregation does not create superior knowledge. Suppose that both types of consumers are uninformed of their types under some market states and are informed under others. In this case, data collection on the noisy perceptions is still informative for the firm, but it does not enable the firm to learn beyond uninformed consumers about their true type. Therefore, consumers do not need to suspect that the firm can take advantage of their uninformed preferences by inducing them to overpay. Results from such a setting are relatively intuitive given what we know about second-degree price discrimination. We summarize the results in Lemma 1.

**Lemma 1.** *Suppose data aggregation does not create superior knowledge. Then the firm is always better off from data collection. However, data collection is never a Pareto improvement because some consumers are strictly worse off.*

As will become subsequently clear, data collection may become a Pareto improvement when the firm obtains superior knowledge. Therefore, it is important to identify the conditions under which superior knowledge is possible from data aggregation.

The first requirement for superior knowledge is that consumers' perceptual errors are highly correlated. Consider, for example, when the external noises are independent for each consumer (i.e., $m_i \perp m_j$); then the firm's knowledge of consumer $i$'s data is irrelevant for consumer $j$. By contrast, when consumers' noisy perceptions are correlated with each other, data aggregation may enable the firm to learn more information than any uninformed consumer who observes only her own data.

A stylized way to implement the correlations across the perceptual noise terms is to assume they are driven by a common market factor, which we term the *market state*. Specifically, we assume that $m \equiv m_L = m_H$. The market state $m \in (0,1)$ is a random variable that affects both types of consumers' initial assessments (not intrinsic values). For instance, years of uninspiring innovations in office furniture—a pessimistic market state—may affect both consumer types in their assessments.

**Assumption 1.** *The market state is the same across consumer types: $m \equiv m_L = m_H$.*

The second requirement for superior knowledge is that the firm can perfectly deduce the market state from the aggregate data set. Denote the data set as $\boldsymbol{\Theta}(m) \equiv \{\theta_H, \theta_L\}$. This requirement implies that the mapping from the market state $m$ to the consumer data set $\boldsymbol{\Theta}(m)$ is injective (i.e., one-to-one). Formally, for any $m$, $m' \in (0,1)$, if $\boldsymbol{\Theta}(m) = \boldsymbol{\Theta}(m')$, then $m = m'$. This requirement thus guarantees the existence of an injective mapping from $\boldsymbol{\Theta}(m)$ to $m$ under all market states. The injective requirement can be reflected in a tractable format with Assumption 2, which states that weak (strong) market states lead to weak (strong) assessments.[6]

**Assumption 2.** *Consumer data is strictly increasing with the market state for any consumer type: $\frac{\partial}{\partial m}\theta_i(m) > 0$ for every $i \in \{H, L\}$ and $m \in (0,1)$.*

The third requirement for superior knowledge is that the firm can perfectly infer the consumer type by deducing the market state. This requirement implies that given any market state $m$, the mapping from the consumer type to the consumer data is injective. Formally, for any $m$, if $\theta_i(m) = \theta_j(m)$, then $i = j$. This requirement thus guarantees that both types of consumers observe different signals under the same market state. Without loss of generality, we assume that the consumer data preserve the same order as consumer values (i.e., $\theta_H(m) > \theta_L(m)$ for any $m$). In this way, the firm can make inferences about the consumer type by observing the rank of the consumer signal in any market state.

**Assumption 3.** *The H-types observe a higher value than the L-types do under the same market state: $\theta_H(m) > \theta_L(m)$ for any $m \in (0,1)$, $i \in \{H, L\}$.*

As will become subsequently clear, Assumption 3 also ensures that there exists a separating equilibrium under any market state $m$.[7]

In summary, under Assumption 1, consumers' perceptual errors are perfectly correlated and influenced by a market state. Under Assumption 2, the firm can perfectly deduce the market state by data aggregation. And under Assumption 3, the firm can infer any uninformed consumer's type by deducing the market state. In combination, these three assumptions ensure that data aggregation always creates superior knowledge for the firm, whenever uninformed consumers exist.

It remains to derive the conditions for the existence of uninformed consumers. Because Assumption 2 ensures that $\theta_i^{-1}$ exists, denote a *transform function*

as $\theta_i^{-1} \circ \theta_j(m)$, which maps from the true market state, $m$, to an inferred market state, $m'$. Under $m'$, the $i$-type would observe the same signal as $j$-type would observe under $m$ (i.e., $\theta_i(m') = \theta_j(m)$). By Definition 1 and Assumption 2, $H$-types are uninformed if and only if $m \in (0, \theta_H^{-1} \circ \theta_L(1))$, and $L$-types are uninformed if and only if $m \in (\theta_L^{-1} \circ \theta_H(0), 1)$. For these intervals to be feasible, we must have $\theta_H^{-1} \circ \theta_L(1) > 0$ and $\theta_L^{-1} \circ \theta_H(0) < 1$, both of which are equivalent to $\theta_H(0) < \theta_L(1)$. The above discussions establish Lemma 2.

**Lemma 2.** *If $\theta_H(0) < \theta_L(1)$, then there exists a market state in which some consumers are uninformed of their type. Furthermore, aggregate consumer data $\{\theta_H, \theta_L\}$ give the firm superior knowledge under any $m \in (0, \theta_H^{-1} \circ \theta_L(1)) \cup (\theta_L^{-1} \circ \theta_H(0), 1)$.*

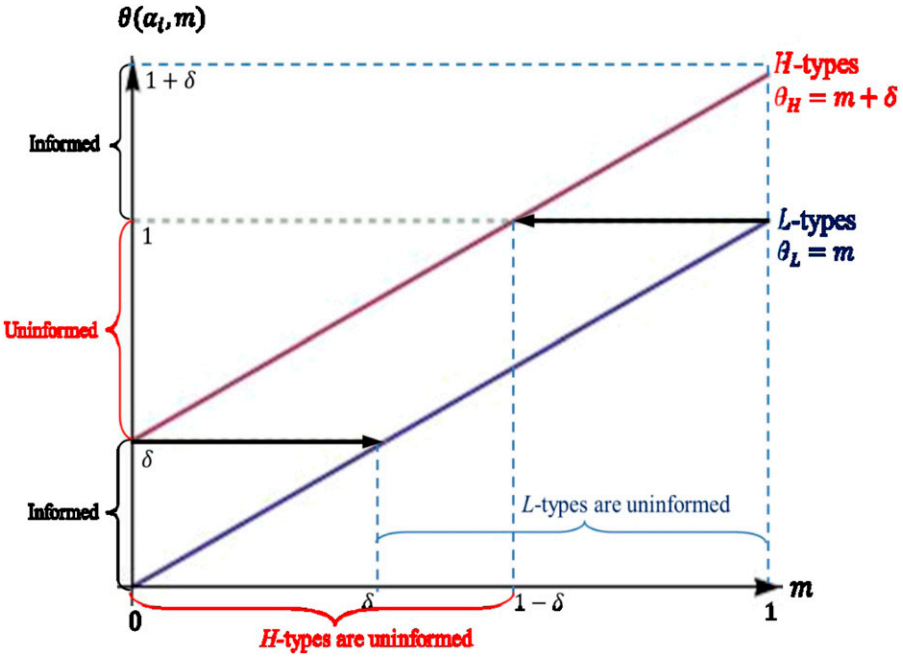
In the remaining sections, we restrict the parameter space so that uninformed consumers may exist (i.e., $\theta_H(0) < \theta_L(1)$). This restriction thus guarantees a positive probability that aggregate consumer data generate superior knowledge for the firm. It is important to bear in mind that the firm's informational advantage does not come from managing sampling or statistical errors but from deducing the market state. Specifically, a single observation $\theta_i \in (\theta_H(0), \theta_L(1))$ is insufficient to solve for $m$. By contrast, the firm can always deduce $m$ by aggregating across types $\{\theta_L, \theta_H\}$. Knowledge of $m$ is important for product line design, because it tells the firm which consumers are uninformed of their types. We illustrate this result by specifying the $\theta$'s as linear functions.

Figure 1 graphically depicts the mechanism described in Lemma 1 for linear forms of $\theta_i(m)$. The model and its equilibrium under generic nonlinear specifications are described in the online appendix. In this figure, we consider the simplest case that satisfies Assumptions 1, 2, and 3. The upper curve represents $\theta_H(m) = m + \delta$, where $\delta > 0$ is a constant, and the lower curve represents $\theta_L(m) = m$. In this case, consumers are uninformed if and only if they observe $\theta \in (\delta, 1)$. The condition of Lemma 2 simply guarantees an interval of $\theta$'s that is potentially received by both types of consumers; that is, $\delta < 1$. Because $\theta_H^{-1} \circ \theta_L(1) = 1 - \delta$, and $\theta_L^{-1} \circ \theta_H(0) = \delta$, $H$-types are uninformed if and only if $m \in (0, 1 - \delta)$, and $L$-types are uninformed if and only if $m \in (\delta, 1)$. Because under any $m$ the firm observes both $m + \delta$ and $m$, it deduces $m$ and learns whether $H$-types and/or $L$-types are uninformed.

### 3.2. Perceptual Errors vs. Sampling Errors

Although we abstract away from the statistical problem, it may be helpful to briefly contextualize the role of data aggregation in the familiar market research contexts. Suppose the firm conducts a conjoint study across a random sample of $n$ consumers in the market. Let $\tilde{\Theta}(\alpha_i, m, \eta_j)$ represent the observed distribution of all consumers' part-worth, where $\alpha_i$ is consumer type $i$'s intrinsic part-worth, $m$ is the common perceptual errors (Assumption 1), and $\eta_j$ is the sampling errors for consumer $j$ of type $i$, independently and identically distributed. Marketers can partition the aggregate observations into clusters according to the known distribution of $\alpha_i$.[8] For example, in our model, the firm can divide the observations into two clusters, one with the smallest $(1 - \lambda)n$ values of $\tilde{\theta}_j$ and the other containing the remainder of the sample (Assumption 3).

**Figure 1.** (Color online) Consumer Data ($\theta$) and Market State ($m$)



Let the average from the $i$th cluster be $\bar{\theta}_i$. Our model is thus relevant when two conditions hold: (1) the firm can estimate cluster means by controlling the sampling errors $\eta_j$, and (2) the firm can estimate the consumers' perceptual errors $m$ from the cluster averages. The first condition holds when statistical methods can unbiasedly estimate $\theta_i$ with $\bar{\theta}_i$ by minimizing the statistical noises from the sampling errors $\eta_j$. The second condition is possible under Assumptions 1, 2, and 3.

By specifying the functional form of $\bar{\boldsymbol{\Theta}}$, the firm can then infer perceptual errors by calculating $\hat{m} = \boldsymbol{\Theta}^{-1}(\bar{\theta}_H, \bar{\theta}_L)$. In our simplified setting without $\eta_j$, the firm needs a sample of only two observations[9] to derive $\hat{m}$, because the mapping from $m$ to $\boldsymbol{\Theta}$ is injective. But a sample size $n > 2$ determines the statistical precision of the estimate $\hat{m}$ by the law of large numbers, when there exist sampling errors.

In summary, even when sampling errors $\eta_j$ exist, data aggregation may enable the firm to acquire private knowledge of consumers' perceptual errors (market state) and, therefore, an understanding of uninformed consumers' beliefs of their type. This understanding is critical when designing the product line, because it helps the firm manage consumers' suspicions of buying the wrong product, ex post. For instance, when $L$-types are uninformed, their expected value for quality exceeds their intrinsic value ($E[\alpha|\theta_L] > \alpha_L$) and they risk overpaying by choosing a product with too-high quality. Similarly, when $H$-types are uninformed, they undervalue the higher-end product ($E[\alpha|\theta_H] < \alpha_H$) and risk buying an inferior product. By not accounting for consumer suspicions, the firm's product line will be suboptimal, as we illustrate in Section 4.

## 4. Consumer Suspicions: A Numerical Example

To illustrate the basic implications of consumer suspicions on product line design, we present a simple numerical example of the above model. Specifically, we show that ignoring the suspicions is unprofitable. As we will see, rational consumers expect to be exploited by the firm's superior knowledge and consequently update their beliefs upon witnessing the chosen product line. We first consider a product line strategy that ignores consumers' updates and then show how a different product line can alleviate consumers' suspicions and convince uninformed $H$-types to raise their valuations.

Suppose that $\alpha_H = 1, \alpha_L = 0.6, \lambda = 0.5, \theta_H = m + 0.5$, and $\theta_L = m$. Because $\theta_i$ is increasing on $m$, the linear functions satisfy Assumptions 2 and 3 and thus enable superior knowledge for the firm. Any consumer who observes $\theta_i(m) \in (0.5, 1)$ is uninformed. Therefore, $H$-types are uninformed of their type when $m \in (0,0.5)$, and $L$-types are uninformed when $m \in (0.5,1)$.

The uninformed consumer's conditional estimate of $\alpha$ is

$$E[\alpha|\theta_i \in (0.5, 1)] = \alpha_L + \frac{\lambda(0.5-0)(\alpha_H - \alpha_L)}{\lambda(0.5-0) + (1-\lambda)(1-0.5)} = 0.8.$$

We summarize the consumers' updated estimate $E[\alpha|\theta_i(m)]$ in Table 1. From Table 1, the market state can be partitioned into three segments. When $m = 0.5$, both types are informed, so it reduces to the standard case (Mussa and Rosen 1978). But when $m \in M_0$ or $M_1$, either $L$-types or $H$-types are uninformed, respectively. By contrast, because $\{\theta_H, \theta_L\}$ differs by $M_0$ and $M_1$, the firm can deduce which consumer segment is uninformed in different market state. For example, when the firm observes the data {0.6,0.8}, it infers that $m \in M_1$ and therefore can deduce that $H$-types are uninformed and underestimating their value.

We start with a naïve approach to the firm's product line design—one that does not consider consumer suspicions. Suppose the firm designs a product line given the distribution of consumers' estimated values in Table 1. As we demonstrate, this product line cannot survive in equilibrium, when rational consumers are suspicious that the firm may exploit its superior knowledge. We establish this claim in the following thought experiment.

First, we examine whether the standard product line design can be a separating strategy. Suppose that when $m \in M_0$, the firm designs a product line $\Psi^{M_0}$ given the consumer beliefs of $\hat{\alpha}_H = 1$ and $\hat{\alpha}_L = 0.8$. Then by Mussa and Rosen (1978) and Moorthy (1984), the optimal product line is $\Psi^{M_0} \equiv \{q_1^{M_0} = 1; p_1^{M_0} = 0.88; q_2^{M_0} = 0.6; p_2^{M_0} = 0.48\}$. Similarly, when $m \in M_1$, the firm designs $\Psi^{M_1}$ given the consumer beliefs of $\hat{\alpha}_H = 0.8$ and $\hat{\alpha}_L = 0.6$. The optimal product line is then $\Psi^{M_1} \equiv \{q_1^{M_1} = 0.8; p_1^{M_1} = 0.56; q_2^{M_1} = 0.4; p_2^{M_1} = 0.24\}$. Because $\Psi^{M_0} \neq \Psi^{M_1}$, any uninformed consumer should correctly infer her type from the separating product line. In particular, an uninformed $L$-type learns $m \in M_0$ whenever she observes the product line $\Psi^{M_0}$ and subsequently infer her value as $\hat{\alpha}_L = 0.6$. Consequently, no one purchases the low-end product in $\Psi^{M_0}$. In fact, the firm would be more profitable in this market state treating $L$-types as if they were informed

**Table 1.** Distribution of Consumers' Estimates for Each Market State

| | Consumer's estimates $\hat{\alpha}_i = E[\alpha|\theta_i(m)]$ | |
|---|---|---|
| Market state | $L$-types | $H$-types |
| $m \in (0.5, 1) \equiv M_0$ | 0.8 (uninformed) | 1 |
| $m \in \{0.5\}$ | 0.6 | 1 |
| $m \in (0, 0.5) \equiv M_1$ | 0.6 | 0.8 (uninformed) |

(i.e., the standard product line with $\hat{\alpha}_H = 1$ and $\hat{\alpha}_L = 0.6$). Therefore, it is not optimal for the firm to implement the standard solution to the product line problem, ignoring consumers' belief updates.

It is then natural to ask whether the firm should implement a pooling strategy to induce the uninformed consumers to purchase the high-end product. For instance, suppose the firm were to implement $\Psi^{M_1}$ for all $m \in M_0 \cup M_1$ and ask whether it can exploit the uninformed $L$-types' overestimation under $M_0$. We now argue that this is impossible under the intuitive criterion (Banks and Sobel 1987, Cho and Kreps 1987). Suppose that $\Psi^{M_1}$ is part of a pooling equilibrium and the uninformed consumers always purchase the high-end product $q_1^{M_1}$. Then the firm profits $\left(0.56 - \frac{0.8^2}{2}\right)\frac{1}{2} + \left(0.24 - \frac{0.4^2}{2}\right)\frac{1}{2} = 0.20$ under $M_1$ and $0.56 - \frac{0.8^2}{2} = 0.24$ under $M_0$. But there exists an out-of-equilibrium product line, $\Psi' \equiv \{q_1' = 1; p_1' = 0.74; q_2' = 0.6; p_2' = 0.36\}$. Under $M_1$, the deviation profit, $\left(0.74 - \frac{1^2}{2}\right)\frac{1}{2} + \left(0.36 - \frac{0.6^2}{2}\right)\frac{1}{2} = 0.21$, is greater than the equilibrium profit of 0.20, given that uninformed consumers purchase $q_1'$. Under $M_0$, however, the incentive of deviation is absent, because the firm profits only $0.74 - \frac{1^2}{2} = 0.24$. Therefore, $\Psi'$ is a profitable deviation only under $M_1$. By the intuitive criterion, any uninformed consumers upon observing $\Psi^{M_1}$ should update their belief that $m \in M_0$ and assess their value as $\alpha_L$. Consequently, no one purchases the high-end product in $\Psi^{M_1}$ under $M_1$. Thus, the pooling strategy fails the intuitive criterion and is suboptimal for the firm.

As the above indicates, neither product line $\Psi^{M_0}$ nor $\Psi^{M_1}$ is optimal for the firm despite being optimal in the standard model, because neither alleviates consumers' suspicions of being tricked into overbuying. To see a product line strategy that overcomes this challenge, first consider the product line $\Psi^{M_0^*} \equiv \{q_1^{M_0^*} = 1; p_1^{M_0^*} = 0.92; q_2^{M_0^*} = 0.2; p_2^{M_0^*} = 0.12\}$, which is identical to the standard case where $\hat{\alpha}_H = 1$ and $\hat{\alpha}_L = 0.6$. Suppose that $m \in M_0$, so then the uninformed $L$-types would purchase $q_2^{M_0^*}$ because the firm is not exploiting their overestimation. In this case, the firm earns $\frac{1}{2}\left(0.92 - \frac{1^2}{2}\right) + \frac{1}{2}\left(0.12 - \frac{0.2^2}{2}\right) = 0.26$. Now suppose that $m \in M_1$. If the firm offers $\Psi^{M_1^*} \equiv \{q_1^{M_1^*} = 1; p_1^{M_1^*} = 0.76; q_2^{M_1^*} = 0.6; p_2^{M_1^*} = 0.36\}$, then the uninformed $H$-types are convinced that $m \in M_1$ as the incentive of offering $\Psi^{M_1^*}$ is absent under $M_0$ relative to using $\Psi^{M_0^*}$, because the firm would earn no greater than $\left(0.76 - \frac{1^2}{2}\right) = 0.26$. Therefore, $\Psi^{M_1^*}$ convinces the uninformed $H$-types of their type and enables the firm to earn $\left(0.76 - \frac{1^2}{2}\right)\frac{1}{2} + \left(0.36 - \frac{0.6^2}{2}\right)\frac{1}{2} = 0.22$. Note that consumer suspicion even improves the firm's profit under $M_1$ (compared with the profit of 0.20 using $\Psi^{M_1}$ in the standard case).

Before generalizing the above result, we use this numerical example to further illustrate how the equilibrium product line with superior knowledge, described above, differs from the standard solution (Mussa and Rosen 1978). The key distinction is that the quality of $L$-types' product, $q_2^{M_1^*} = 0.6$, is not downward-distorted. Rather, the equilibrium quality of the low-end product is economically efficient because the marginal cost of quality equals the marginal value: $q_2^{M_1^*} = \alpha_L$. This is because the firm reduces the high-end product's price to convince uninformed $H$-types of their type. Consequently, $H$-types have less incentive to choose the cheaper alternative, and the conventional cannibalization constraint may not bind. Note that the firm's margin from an $L$-type is 0.18 with $\Psi^{M_1^*}$ and 0.10 with $\Psi^{M_0^*}$. Therefore, the relaxation of the cannibalization constraint under $M_1$ enables the firm to increase the profit margin from the $L$-types by improving the low-end product's quality, thus restoring the economic efficiency in the product line.

The above example illustrates how the firm resolves the central challenge of designing a product line with superior knowledge. In contrast to the classic product line design problem, the firm must consider how the uninformed consumers may update their beliefs after observing the products. In addition, the consumers' purchase decisions must assess both the firm's profit maximization and its estimates of their updated beliefs. This strategic interaction implies signaling costs, which are not present in the classic product line design problem, with benefits accruing to consumers. Therefore, as we demonstrate in Sections 5 and 6, superior knowledge afforded by data collection has surprising implications for consumer welfare.

## 5. Equilibrium

The numerical example demonstrated a special case in which at least one consumer is informed under every market state. We maintain this restriction by assuming $\theta_H^{-1} \circ \theta_L(1) \le \theta_L^{-1} \circ \theta_H(0)$. But our equilibrium results generalize to the case in which both types of consumers are uninformed under the same market state, $m \in (\theta_L^{-1} \circ \theta_H(0), \theta_H^{-1} \circ \theta_L(1))$ (see the online appendix).

In the basic model, we partition the market space by two segments: $\mathcal{M}^{\mathbf{0}} \equiv [\theta_L^{-1} \circ \theta_H(0), 1)$, where $H$-types are informed, and $\mathcal{M}^{\mathbf{1}} \equiv (0, \theta_H^{-1} \circ \theta_L(1))$, where $H$-types are uninformed. Denote $m_H$ ($m_L$) as an uninformed consumer's inferred market state given she is an $H$-type ($L$-type). Lemma B.1 in the appendix ensures that for every uninformed consumer, $m_H \in \mathcal{M}^1$ and $m_L \in \mathcal{M}^0$.

Because the firm's product line design may change the uninformed consumers' beliefs and purchase decisions, it implies a signaling game for which we employ the concept of a perfect Bayesian equilibrium (PBE). As in many signaling games, there is a multitude of equilibria, most of which are supported by unreasonable out-of-equilibrium consumer beliefs. We employ the intuitive criterion (Cho and Kreps 1987) to eliminate these equilibria. Our equilibrium also survives a stronger version of belief refinement, the D1 criterion (Banks and Sobel 1987, Cho and Sobel 1990). In particular, upon observing a deviation, uninformed consumers assign zero probability to a market state, in which whenever the firm has weak incentives to deviate, it has strong incentives to deviate in another state. Furthermore, if there exists multiple perfect Bayesian equilibria with the same firm profit, then we select those with the Pareto-efficient outcome.

We first introduce some notations. Denote the uninformed consumer $i$'s purchase decision as $r_i \in$ {Purchase product $k$, No purchase} and the firm's product line design as $\psi$. Let $\mu_i \equiv \mu(\psi)$ be consumer $i$'s estimated probability that she is an $H$-type when observing $\psi$, and let$(\psi^*, r_i^*, \mu_i)_{i=H,L}$ be a PBE, which specifies the equilibrium product line, consumer choice, and belief. Denote $\psi'$ as the firm's deviation strategy, $\mathrm{BR}(\psi', \mu_i)$ as consumer $i$'s best response (purchase decision) given belief $\mu(\psi')$,[10] and $\hat{\prod}(m, r_i, \psi')$ as her calculation of the firm's deviation profit under the $m$ state.

Because the firm's product line design influences the consumers' belief update, it incurs an additional constraint in the optimization problem. For example, under any market state in $\mathcal{M}^1$, $H$-types are uninformed. Thus, to ensure that $H$-types choose the high-end product,[11] the firm must convince them of their type. Therefore, the optimal design problem under $\mathcal{M}^1$ requires an additional constraint (iii) on how the observed product line may induce the consumers to update their beliefs:

$$\max_{\{p_1,p_2,q_1,q_2\geq 0\}} \lambda\left(p_1-\frac{q_1^2}{2}\right)+(1-\lambda)\left(p_2-\frac{q_2^2}{2}\right)$$

subject to

1. $[\mu_H\alpha_H+(1-\mu_H)\alpha_L]q_1-p_1\geq \max\{[\mu_H\alpha_H+(1-\mu_H)\alpha_L]q_2-p_2,0\}$;
2. $\alpha_L q_2-p_2\geq \max\{\alpha_L q_1-p_1,0\}$;
3. $\mu_H$ is sequentially rational, Bayesian whenever possible, and survives both the intuitive and D1 criteria (Banks and Sobel 1987, Cho and Kreps 1987, Cho and Sobel 1990).[12]

Proposition 1 characterizes the equilibrium product line and beliefs under every market state.

**Proposition 1.** *There exists a unique perfect Bayesian equilibrium $(\psi^*, r_i^*, \mu_i)_{i=H,L}$ such that the following properties hold ($\psi$ is characterized in Table 2):*

(i) *The equilibrium product line is separating by $\mathcal{M}^0$ and $\mathcal{M}^1$.*

(ii) *In any market state $m$, $H$-types purchase the high-end Product 1; $L$-types either purchase the low-end Product 2 or make no purchases.*

(iii) *Consumer $i$ updates her belief upon observing any product line $\psi$ as follows: $\mu_i=1$, if and only if $\hat{\Pi}(\mathcal{M}^0, r_i, \psi)\leq \Pi^*(\mathcal{M}^0)$, where $r_i\in \mathrm{BR}(\psi,\mu_i=1)$; otherwise, $\mu_i=0$.*

(iv) *The equilibrium uniquely survives the intuitive criterion and the D1 criterion.*

(v) *The equilibrium is Pareto efficient.*

Proposition 1 implies that any pooling equilibrium fails the intuitive criterion (IC). This result generalizes the discussion in Section 4 for the numerical example. The intuition is that for any pooling equilibrium that may trick the uninformed $L$-types into overpaying for the $H$-product, there always exists

**Table 2.** The Equilibrium Product Line Design with Data Collection

| | | | $\alpha_L\leq\lambda\alpha_H$ | | $\alpha_L>\lambda\alpha_H$ | |
|---|---|---|---|---|---|---|
| $m$ | $\psi^*(m)$ | | $\lambda\leq\frac{1}{2}$ | $\lambda>\frac{1}{2}$ | $\lambda\leq\frac{1}{2}$ | $\lambda>\frac{1}{2}$ |
| $m\in\mathcal{M}^0$ | Informed $H$-types | $q_1^*(\mathcal{M}^0)$ | $\alpha_H$ | | $\alpha_H$ | |
| | | $p_1^*(\mathcal{M}^0)$ | $\alpha_H^2$ | | $\frac{(\alpha_H-\alpha_L)^2}{1-\lambda}+\alpha_H\alpha_L$ | |
| | Uninformed $L$-types | $q_2^*(\mathcal{M}^0)$ | 0 | | $\frac{\alpha_L-\lambda\alpha_H}{1-\lambda}$ | |
| | | $p_2^*(\mathcal{M}^0)$ | 0 | | $\frac{\alpha_L^2-\lambda\alpha_H\alpha_L}{1-\lambda}$ | |
| $m\in\mathcal{M}^1$ | Uninformed $H$-types | $q_1^*(\mathcal{M}^1)$ | $\alpha_H$ | $\alpha_H$ | $\alpha_H$ | $\alpha_H$ |
| | | $p_1^*(\mathcal{M}^1)$ | $\frac{(1+\lambda)\alpha_H^2}{2}$ | $\frac{(1+\lambda)\alpha_H^2}{2}$ | $\frac{\alpha_H^2-2\lambda\alpha_H\alpha_L+\alpha_L^2}{2(1-\lambda)}$ | $\frac{\alpha_H^2-2\lambda\alpha_H\alpha_L+\alpha_L^2}{2(1-\lambda)}$ |
| | Informed $L$-types | $q_2^*(\mathcal{M}^1)$ | $\alpha_L$ | $\frac{(1-\lambda)\alpha_H^2}{2(\alpha_H-\alpha_L)}$ | $\alpha_L$ | $\frac{(1-2\lambda)\alpha_H+\alpha_L}{2(1-\lambda)}$ |
| | | $p_2^*(\mathcal{M}^1)$ | $\alpha_L^2$ | $\frac{(1-\lambda)\alpha_H^2\alpha_L}{2(\alpha_H-\alpha_L)}$ | $\alpha_L^2$ | $\frac{(1-2\lambda)\alpha_H\alpha_L+\alpha_L^2}{2(1-\lambda)}$ |

an out-of-equilibrium strategy, in which the deviation is profitable under $\mathcal{M}^1$ but not under $\mathcal{M}^0$. Therefore, any uninformed consumer with an IC belief should infer that $m \in \mathcal{M}^0$ and thus reject the $H$-product. Proposition 1 then establishes that only a separating equilibrium may survive IC. Formal details are relegated to Step 3.2 of the proof in the appendix.

The equilibrium belief specified in Proposition 1(iii) can be intuitively expressed in the following statement: "I want to determine why the firm has chosen to implement this product line. If the firm has an incentive to deviate in $\mathcal{M}^0$ (I am an $L$-type) when I behave as an $H$-type, then I should assess that I am certainly an $L$-type." This belief survives not only IC but also D1. The single-crossing property is implied because the firm profits more by upselling to the uninformed $L$-types than to the $H$-types. Therefore, whenever the firm has a weak incentive to deviate in $\mathcal{M}^1$, it has a strong incentive to deviate in $\mathcal{M}^0$. Because this belief ensures that uninformed $L$-types learn their type under $\mathcal{M}^0$, by Proposition 1(i), the threshold, $\Pi^*(\mathcal{M}^0)$, equals $\Pi_F^*$, the firm's profit in the standard, full-information case derived in Mussa and Rosen (1978) and Moorthy (1984). Therefore, $\Pi_F^*$ is a key threshold for consumers when they form beliefs about what the firm can earn in off-equilibrium paths. Thus, consumers use $\Pi_F^*$ to infer the firm's incentives of deviation under $\mathcal{M}^0$. If a deviation is possible, the consumers should assign a higher probability to the $L$-types (under $\mathcal{M}^0$). By applying this probability assignment iteratively, the uninformed consumers assess that they are the $L$-types with probability 1.

We can interpret the uninformed consumer's equilibrium belief as maintaining a level of suspicion that prevents the firm from overcharging her. That is, the firm cannot mislead $L$-types to overbuy because $\alpha_L q_2^*(\mathcal{M}^0) - p_2^*(\mathcal{M}^0) = 0$. To interpret this result, suppose that the firm could offer a product line $\psi'$ under $\mathcal{M}^0$ to convince an uninformed $L$-type that she is an $H$-type. Then by Proposition 1(iii), it is required that $\hat{\Pi}(\mathcal{M}^0, r_i, \Psi') \le \Pi^*(\mathcal{M}^0)$. Therefore, $L$-types' suspicions are sufficient to make any deceptive strategy suboptimal for the firm, so the firm with superior knowledge cannot exploit uninformed $L$-types' perceptual errors. Note that for those market states under $\mathcal{M}^0$, the firm may prefer to skip $L$-types to mitigate cannibalization by setting $q_2^*(\mathcal{M}^0) = 0$. The conditions of partially market coverage $\left(\alpha_L \le \lambda\alpha_H \text{ and } \lambda \le \frac{1}{2}\right)$ are consistent with the classic results.

When $m \in \mathcal{M}^1$, however, superior knowledge over uninformed $H$-types forces the firm to offer a different product menu. The firm would inform any $H$-type of her intrinsic value to increase the profit margin by improving the high-end product's quality. Therefore, the firm prefers to signal the state $\mathcal{M}^1$ to uninformed $H$-types. By Proposition 1(iii), this requires that the profit margin of the high-end product does not exceed a threshold. Specifically, the firm's product line optimization in $\mathcal{M}^1$ is

$$\max_{\{p_1,p_2,q_1,q_2\ge 0\}} \lambda\left(p_1 - \frac{q_1^2}{2}\right) + (1-\lambda)\left(p_2 - \frac{q_2^2}{2}\right)$$

subject to

$(\mathrm{IC_H})\ \alpha_H q_1 - p_1 \ge \alpha_H q_2 - p_2,$
$(\mathrm{IR_H})\ \alpha_H q_1 - p_1 \ge 0,$
$(\mathrm{IC_L})\ \alpha_L q_2 - p_2 \ge \alpha_L q_1 - p_1,$
$(\mathrm{IR_L})\ \alpha_L q_2 - p_2 \ge 0,$
$(\mathrm{SC})\ p_1 - \frac{q_1^2}{2} \le \Pi_F^*.$

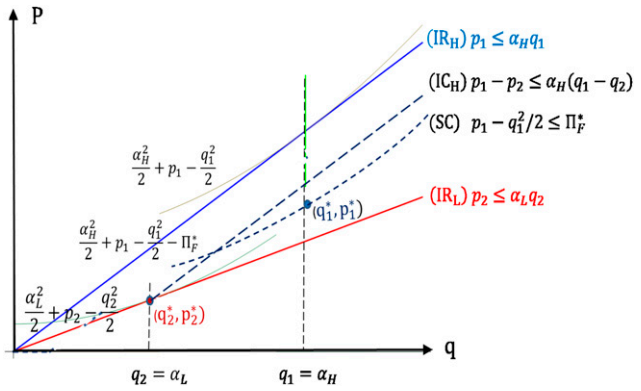
Like in the standard model, the participation constraint ($\mathrm{IR_L}$) is always binding, but ($\mathrm{IR_H}$) and ($\mathrm{IC_L}$) are not. But unlike in the standard model, the incentive compatibility constraint ($\mathrm{IC_H}$) may no longer bind as a result of the signaling constraint (SC). Note that the firm will set $p_1^*$ sufficiently low rather than reducing $q_1^*$, because a low $q_1^*$ may strengthen the cannibalization constraint ($\mathrm{IC_H}$) and reduce the profit margin from serving the informed $L$-types. In addition, it is Pareto optimal to set the quality $q_1^*$ at the efficient level and lower the price $p_1^*$, given the same profit margin from $H$-types. Consequently, the firm raises $q_2^*$ above the standard distorted level to charge $L$-types a higher price. Putting it differently, the burden of signaling to $H$-types may reverse the quality distortion and restore efficiency to the product line. Proposition 2 formalizes this finding.

**Proposition 2.** *Compared with the standard model, consumers obtain higher-quality products on average when the firm acquires superior knowledge. In addition, the efficiency of product line design is fully restored when H-types are uninformed and sufficiently small in population; that is,* $q_1^*(\mathcal{M}^1) = \alpha_H$ *and* $q_2^*(\mathcal{M}^1) = \alpha_L$ *if and only if* $\lambda \le \frac{1}{2}$.

The intuition of Proposition 2 is that when the signaling constraint is too strong, the firm must reduce the price $p_1$ to convince the uninformed $H$-types of their value. As $H$-types have less incentive to choose the cheaper alternative, the cannibalization constraint is relaxed. Because the firm has incentives to raise $q_2$ for a higher profit margin from $L$-types, the product design may restore its economic efficiency and thus maximize the total social welfare. This scenario is most likely when $\lambda$ is sufficiently low, because then (1) the proportion of profit from serving $L$-types relative to serving $H$-types is higher, and (2) the cannibalization constraint is more likely dominated by the signaling constraint.

Figure 2 illustrates the intuition in Proposition 2. The two U-shaped curves are the firm's iso- margin curves for each type of consumers. The dashed curve

**Figure 2.** (Color online) The Equilibrium Product Line Under $\mathcal{M}^1$




defines the signaling constraint (SC), $p_1 - \frac{q_1^2}{2} \le \Pi_F^*$, a condition that is absent in the standard model (Mussa and Rosen 1978). The two solid lines define the participation constraints ($IR_H$ and $IR_L$), above which consumers prefer not to purchase. The dashed line specifies the cannibalization constraint ($IC_H$), which is equivalent to $\alpha_H q_1 - p_1 \ge \alpha_H q_2 - p_2$. The slope between the two product points $(q_1, p_1)$ and $(q_2, p_2)$ must be lower than the dashed line $IC_H$; otherwise, the informed $H$-types would prefer the cheaper alternative (the low-end product). The signaling constraint, relevant whenever $H$-types are uninformed, forces the firm to lower the profit margin from $H$-types. If the signaling cost is large enough, the vertical line $q = \alpha_H$ intercepts with the dashed curve SC lower than with the dashed line $IC_H$. In this case, the signaling constraint is stronger than the cannibalization constraint, because the slope between the two equilibrium points $(q_1^*, p_1^*)$ and $(q_2^*, p_2^*)$ is lower than the dashed line $IC_H$. Therefore, the cannibalization constraint $IC_H$ is not binding, the firm will upgrade the low-end product to $(q_2^*, p_2^*)$, and the efficiency of the product line is restored.

The condition on $\lambda$ to obtain the efficient product line design is simply that the firm finds it profitable to raise the profit margin serving $L$-types. The effort to signal the uninformed $H$-types implies a "corrective" force against the monopoly distortions of product line cannibalization. This corrective force will help explain the incentives and welfare implications for data collection in Section 6.

Another implication of Propositions 1 and 2 relates to mandatory disclosure laws proposed by privacy advocates (e.g., RECAP, as in the regulation of "Record, Evaluate and Compare Alternative Prices," and Kamenica et al. 2011), in which firms are required to share their collected data with consumers through a trusted third party. Our results suggest that such regulatory disclosure of firm information may reduce total consumer surplus and social welfare when consumers are rational. When consumers are credibly informed by the third party, the firm uses a full-information equilibrium to better exploit $H$-types without the burden of signaling. Therefore, the interference of the third party may reduce consumers' expected surpluses. In addition, because data collection enables the firm to raise the quality of the low-end product closer to the efficient level, the total social welfare without mandatory disclosure is higher relative to the standard case with inefficient product line design. This conclusion holds in our model because consumers are rational when interacting with the firm. If, however, consumers are boundedly rational or unaware of data collection, mandatory data disclosure may have different welfare implications.

## 6. Analyses

In this section, we examine the incentives and the welfare implications of superior knowledge. We start with the benchmark equilibrium when the firm does not collect data. By comparing this equilibrium with that of Proposition 1, we can characterize the conditions for the firm's optimal decision of data collection and the consumers' surpluses. In Section 6.1, we examine whether it is firm optimal to collect data. In Section 6.2, we analyze the ex ante welfare implications. Section 6.3 characterizes the parameter regions in which data collection is ex post Pareto optimal.

Without loss of generality, we restrict the parameter space that $H$-types are uninformed when $m \in (0, \frac{1}{2})$ and $L$-types are uninformed when $m \in (\frac{1}{2}, 1)$. Furthermore, we normalize $\alpha_H = 1$ and denote $\alpha = \alpha_L$. Therefore, $\alpha$ measures the ratio between values, which we call *consumer homogeneity*.

### 6.1. Is Superior Knowledge Always Profitable?

We begin by solving the product line when the firm does not observe consumer data $\{\theta_L, \theta_H\}$. In this scenario, the firm learns less than consumers; thus it uses only the prior knowledge on the market distribution of types to design products. In expectation, the firm anticipates three distinct segments of consumers. A portion of $(\frac{\lambda}{2})$ are informed $H$-types, a portion of $(\frac{1-\lambda}{2})$ are informed $L$-types, and the rest are uninformed. The uninformed consumers estimate their value as $e \equiv \lambda + (1-\lambda)\alpha$, because either type has the same probability to be uninformed.

In this case, the firm's product design problem is then posed as follows:

$$\max_{\{p_1,p_2,p_3,q_1,q_2,q_3\}} \frac{\lambda}{2}\left(p_1 - \frac{q_1^2}{2}\right) + \frac{1}{2}\left(p_2 - \frac{q_2^2}{2}\right) + \frac{(1-\lambda)}{2}\left(p_3 - \frac{q_3^2}{2}\right)$$

subject to $\alpha_j q_j - p_j \geq \max\{\alpha_j q_k - p_k, 0\}$ for every $k, j = 1, 2, 3$, and $k \neq j$, where $\alpha_1 = 1$, $\alpha_2 = e \equiv \lambda + (1-\lambda)\alpha$, $\alpha_3 = \alpha$, and $p_j,\ q_j \geq 0$.

The firm may prefer to skip $L$-types to mitigate cannibalization. Whether it is optimal to serve $L$-types depends on the relative portion of $H$-types ($\lambda$) and the homogeneity between consumers' values ($\alpha$). A small $\alpha$ is interpreted as a large disparity between consumer types. For small levels of consumer homogeneity, $\alpha \leq \frac{\lambda+\lambda^2}{1+\lambda^2}$, the firm has a stronger incentive to extract surplus from the informed $H$-types and uninformed consumers. Thus, it simply abandons its lowest-quality product to keep up the prices for the two higher-end products ($\hat{p}_1$ and $\hat{p}_2$). Denote the equilibrium profit and consumer surplus as $\Pi^*_{nc}$, and $\mathrm{CS}^*_{nc}$, respectively, where the subscript $nc$ refers to not collecting data. Denote the equilibrium product quality and price as $\hat{q}$ and $\hat{p}$, respectively.

Proposition 3 characterizes the equilibrium for the different parametric conditions.

**Proposition 3.** *Suppose the firm does not collect (nc) consumer data. The equilibrium product line strategy $\hat{\Psi}$, firm profits $\Pi^*_{nc}$, and consumer surplus $\mathrm{CS}^*_{nc}$ are given in Table 3.*

Uninformed consumer preferences lead to possibly three segments of consumers, even though there are only two types. The uninformed consumers may obtain a negative surplus. Specifically, $L$-types may overbuy the high-end product and thus be worse off ex post than making no purchases. Furthermore, the firm utilizes the standard distortions of lower-quality products to mitigate cannibalization, which increases both the product line's length (number of products) and degree of differentiation (disparity of quality) relative to the standard case.

By comparing the equilibrium profits from Propositions 1 and 3, we can identify the conditions under which the firm does not collect data. This is the case whenever consumer suspicions are so severe that the cost is relatively higher than the benefit to signal the superior knowledge. Denote the equilibrium profit in Proposition 1 as $\Pi^*_c$, where the subscript $c$ refers to collecting data. Proposition 4 shows the condition when data collection occurs in equilibrium.

**Proposition 4.** *The firm prefers to collect data if and only if $\lambda < (3-\sqrt{5})/2$ or $\alpha < \bar{\alpha}$, where there always exists a unique $\bar{\alpha} \in (0,1)$ that solves $\Pi^*_c(\bar{\alpha}) = \Pi^*_{nc}(\bar{\alpha})$.*

Increasing $\lambda$ and $\alpha$ mean more relative surplus from $H$-types and, therefore, increasing profits regardless of data collection. However, $\Pi^*_{nc}$ increases in these variables at a faster rate than $\Pi^*_c$. With no data collection, an increase in $\lambda$ or $\alpha$ directly raises an uninformed consumer's willingness to pay for the product with quality $\hat{q}_2$. By contrast, with data collection, the signaling cost tempers the increase in firm's profit $\Pi^*_c$.

In summary, although collecting data can be beneficial, it arouses consumers' suspicions that they might be tricked by upselling. The suspicions induce the firm to undertake costly signaling, which cedes information rents to the $H$-types. This signaling cost becomes sufficiently burdensome to the firm when there is a high concentration of value in this segment (i.e., large values of $\lambda$ and $\alpha$). We discuss these and other welfare implications in Section 6.2.

### 6.2. Welfare Implications (Ex Ante)

To assess when superior knowledge benefits the consumers on average, we also compare the equilibrium in Propositions 1 and 3. When data are collected, uninformed $L$-types correctly update their belief and purchase the lower-quality product, and uninformed $H$-types must be convinced by a lower price $p^*_1 < \hat{p}_1$. Therefore, data collection benefits the uninformed consumers ex ante in two ways: (1) $L$-types may avoid overbuys, and (2) $H$-types may obtain higher surplus because of the signaling cost. But because superior knowledge also allows the firm to reduce the length of the product line, the informed $H$-types may obtain lower rent from the relaxed incentive compatibility constraint, especially in the cases of a large $\lambda$ and a small $\alpha$. Consequently, the impact of data collection on consumer surplus is ambiguous without further conditions, which we now explore in Proposition 5.

**Table 3.** The Equilibrium Product Line Design Without Data Collection

| $\hat{\Psi}$ | | $\alpha \in \left(0, \frac{\lambda+\lambda^2}{1+\lambda^2}\right]$ | $\alpha \in \left(\frac{\lambda+\lambda^2}{1+\lambda^2}, 1\right)$ |
|---|---|---|---|
| Informed $H$-types | $\hat{q}_1$ | $1$ | $1$ |
| | $\hat{p}_1$ | $(1-\lambda^2+\lambda^2 e)(1-\alpha)+e\alpha$ | $(1-\lambda^2+\lambda^2 e)+e\alpha-\frac{(1+\lambda^2)(e-\alpha)}{1-\lambda}\left(\alpha-\frac{\lambda+\lambda^2}{1+\lambda^2}\right)$ |
| Uninformed consumers | $\hat{q}_2$ | $\alpha+\lambda^2(1-\alpha)$ | $\alpha+\lambda^2(1-\alpha)$ |
| | $\hat{p}_2$ | $e\alpha+\lambda^2 e(1-\alpha)$ | $e\alpha+\lambda^2 e(1-\alpha)-\frac{(1+\lambda^2)(e-\alpha)}{1-\lambda}\left(\alpha-\frac{\lambda+\lambda^2}{1+\lambda^2}\right)$ |
| Informed $L$-types | $\hat{q}_3$ | $0$ | $\alpha-\frac{\lambda(1+\lambda)}{1-\lambda}(1-\alpha)$ |
| | $\hat{p}_3$ | $0$ | $\alpha^2-\frac{\lambda(1+\lambda)}{1-\lambda}\alpha(1-\alpha)$ |
| Firm's expected profit | $\Pi^*_{nc}$ | $\frac{(1+\lambda)(\lambda(1-e)^2+e^2)}{4}$ | $\frac{(1+\lambda)(\lambda(1-e)^2+e^2)}{4}+\frac{\alpha^2[(\lambda^2+\lambda)(\alpha-1)+(1-\lambda)]^2}{4(1-\lambda)}$ |
| Consumers' expected surplus | $\mathrm{CS}^*_{nc}$ | $(1-\lambda)[\alpha+\lambda^2(1-\alpha)]$ | $\frac{1}{4}(1+\lambda)(\lambda(1-e)^2+e^2)$ |

**Proposition 5.** *Relative to no data collection, data collection with superior knowledge implies that*

(i) *uninformed consumers always obtain higher expected surplus, ex ante; and*

(ii) *the average consumer surplus is lower, ex ante, if and only if both of the following conditions hold:*

$$\lambda \geq \frac{\sqrt{2}}{2} \text{ and } \alpha < \underline{\alpha} \equiv \frac{\sqrt{2\lambda^2 - 1} - 2\lambda^2 + 1}{2(1-\lambda^2)}.$$

As $\alpha$ decreases and $\lambda$ increases, consumers become more heterogeneous, and their prior expectation to be an $H$-type is greater, which implies that $H$-types have higher potential rent from the cannibalization constraint when there is no data collection. In this case, the data-collecting firm also leaves lower surplus to convince $H$-types. Therefore, when $H$-types' loss outweighs $L$-types' gains, data collection reduces consumers' surplus on average.

### 6.3. Pareto Optimality (Ex Post)

It is an immediate consequence of Proposition 5 that data collection may lead to an ex ante Pareto-optimal outcome. That is, under some conditions, average consumers at the start of the game can expect to be better off when their data are collected. However, one can ask the question of whether both types of consumers, rather than on average, are ever simultaneously better off ex post. This stronger condition requires that each type of consumer obtains higher expected surplus in all possible realizations of the equilibrium.

**Proposition 6.** *Suppose the conditions of Proposition 4 hold, so that the firm profits from data collection. Then, relative to no data collection, the equilibrium outcome makes both types of consumers strictly better off ex post, if and only if* $\alpha > \underline{\alpha}$.

Proposition 6 gives the precise conditions for which data collection is a strict Pareto improvement over no data collection. The intuition can be seen by first noting that $L$-types are strictly better off with data collection by avoiding overbuys. Second, uninformed $H$-types may acquire additional surpluses as a result of the signaling cost. Recall that when signaling occurs in equilibrium, it surpasses the incentive compatibility constraint and returns more surplus to $H$-types. In addition, they can avoid underbuys. Third, informed $H$-types may also benefit from data collection by paying a lower price. By contrast, the uninformed firm may underestimate the proportion of informed $H$-type and require a higher profit margin.

The ex post optimality implied in Proposition 6 is created by the improved efficiency of the product line design as a result of the signaling constraint, which counters the incentive compatibility constraint. In the case without data collection, the longer product line with three products induces strong cannibalization concerns, reducing the efficiency of the product line design. In addition, uninformed $L$-types overbuy and obtain a negative surplus. With data collection, however, the consumers know that the firm has private information, and their suspicions of being tricked confound the firm's price discrimination ability, forcing the firm to use the product line as a credible message to help the consumers learn their type. The signaling process also allows the firm to raise the quality of its lower-quality product back toward the efficient level $q_2^*(\mathcal{M}^1) \to \alpha_L$. Therefore, data collection may raise total social welfare, which is subsequently split across all three agents: the firm, $H$-types, and $L$-types. Data collection also corrects the classic inefficiencies (Mussa and Rosen 1978) associated with adverse selection in product line design.

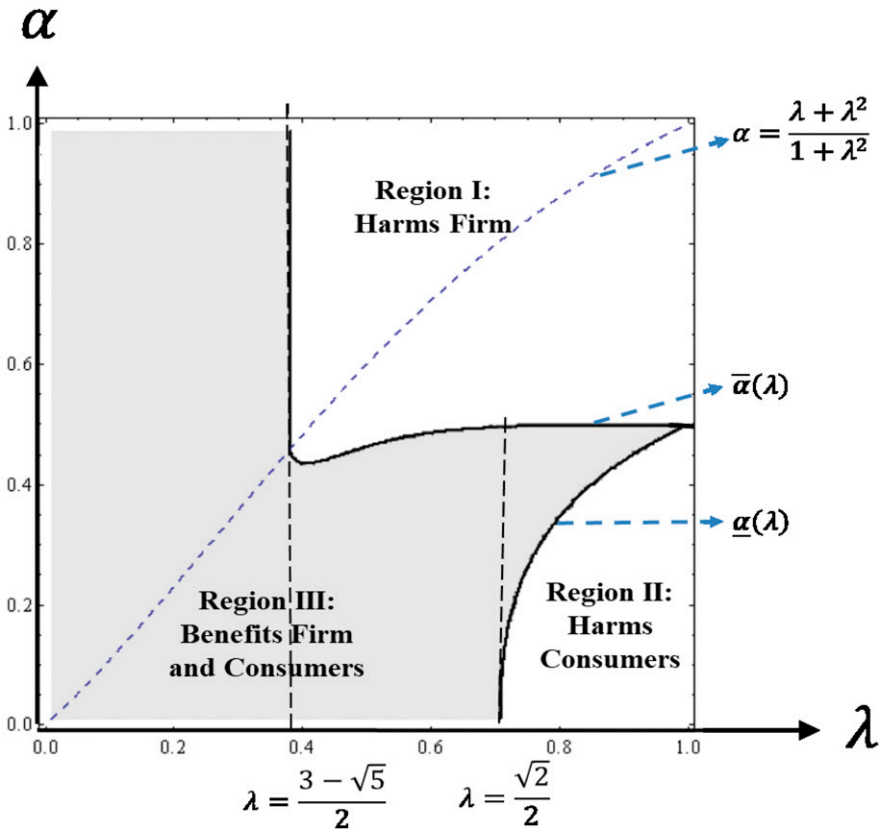
Figure 3 depicts the complementary set of the conditions in Proposition 4 as the upper right region (region I), the conditions in Proposition 5(ii) as region II, and the conditions in Proposition 6 as region III. The vertical axis represents the consumer homogeneity $\alpha \equiv \alpha_L/\alpha_H$, and the horizontal axis represents the population proportion of the $H$-types, $\lambda$. The dashed diagonal curve is the iso-profit boundary along which the firm offers two or three products when it does not collect data from consumers, which is characterized in Proposition 3.

To interpret Figure 3, first consider the region $\alpha \geq \frac{\lambda+\lambda^2}{1+\lambda^2}$. Because $\frac{\lambda+\lambda^2}{1+\lambda^2} > \lambda$, we must have $\alpha \geq \lambda$. Therefore by Propositions 2 and 3, the firm serves both market segments, regardless of whether it collects data or not, across regions I and III. The result implies that the cannibalization concerns are not too severe in either case to be decisive for data collection. Instead, the true incentives for data collection reflect the benefits of overcharging uninformed $L$-types relative to convincing uninformed $H$-types to purchase the higher-quality product. Next, consider $\alpha < \frac{\lambda+\lambda^2}{1+\lambda^2}$. By Proposition 3, cannibalization concerns are so strong in this region that the firm, without data collection, offers only two products (informed $L$-types are not served). The boundary defined by $\bar{\alpha}$ in Proposition 4 is everywhere below the threshold $\frac{\lambda+\lambda^2}{1+\lambda^2}$ but above the threshold $\underline{\alpha}$ in Proposition 5(ii). Suppose the firm is faced with parameter values just below $\underline{\alpha}$; then the firm has an incentive to collect data, but the average consumers are worse off because their signaling benefit cannot compensate for the informed $H$-types' loss in information rent.

## 7. Conclusion

This paper examines the incentives of the product line design when the firm collects data from consumers who are uncertain of their preferences. Such situations, we argue, can arise for innovative products for

**Figure 3.** (Color online) Welfare Implications of Superior Knowledge




which consumers' initial perceptions of prototypes are subject to perceptual errors. We show conditions in which the firm, by collecting data on consumers' perceptions, identifies the extent of perceptual errors and estimates consumer preferences better than the consumers themselves. In this setting, we seek to understand who benefits when a firm has such an informational advantage over consumers.

Our focus on superior knowledge of consumer preferences is intended to capture the modern development of intensive market research and digital technologies. Specifically, our research asks whether superior knowledge leads to exploitation of consumers or better-suited products. When consumers are perfectly informed of their own preference, previous research suggests that collecting individual data generally diminishes consumer surplus, because the firm is better equipped to price discriminate. With uninformed consumer preferences, however, we show that this is not always the case.

This research uncovers a novel mechanism in which data aggregation can create superior knowledge and mutually benefit the firm and every consumer. When consumers' perceptions of preferences are correlated across types, data aggregation enables the firm to learn superior preference information that exceeds consumers' prior knowledge. Although this suggests an informational advantage for the firm, consumers' rational suspicions may confound the firm's ability to exploit its superior knowledge. In fact, to convince consumers of their high value (*H*-types in the model), the firm must reverse the classic product line distortions to convince the uninformed *H*-types to buy the high-end product. As a result, the cannibalization constraint may no longer bind. It is worth noting that despite the additional signaling constraint, the firm may still benefit from collecting data, because it obtains more profit by serving *L*-types products at the efficient level.

By contrast, when there is no data collection (and no superior knowledge), the firm designs the "second-best" product line to extract the usual monopoly rents without raising indirect costs as a result of consumer suspicions. Without data collection, the firm is forced to provide up to three products, which includes a medium-quality product for the uninformed consumers. The longer product line is beneficial for the consumers because the firm must attend to increased cannibalization concerns relative to fewer products.

Our results provoke some conventional wisdom in the debate on consumer data collection. First, a firm's informational advantage over consumer preference data does not necessarily lead to price exploitation, because consumers' rational suspicions prevent them

from being tricked. This also suggests that a firm may have an interest in communicating its knowledge about consumers through its product line, which can correct the classic monopoly distortion in product line design.

Second, this research challenges the meaning of private consumer information. In our model, the aggregation of consumer data can be more informative than the sum of the individual knowledge. With access to data on a broader set of consumers, the firm has the advantage of controlling for the environmental noises that affect individual consumers' perceptions. In particular, data aggregation enables the firm to acquire private information on consumers' imperfect beliefs of their value. This suggests that definitions of private consumer information may be broader than traditionally assumed. Specifically, if data aggregation creates new knowledge, then one could ask whether the new knowledge is the firm's own intellectual property.

Related to the above point is the issue of forced disclosure of a firm's private information. Some researchers have suggested forced disclosure of a firm's consumer data would generally benefit consumers (Kamenica et al. 2011). Our research suggests that this may not always be the case. Whenever the firm obtains superior knowledge on consumer preferences from data collection and consumers know their data have been collected, then consumers' rational suspicions may force the firm to return surpluses via the firm's incentive to communicate to high-valuation consumers. If a credible third party were to communicate the information to consumers directly without the need of convincing, then the firm would be able to implement its standard product line design, which can be worse for consumers because it alleviates the firm's burden of communicating its private information.[13] Our research suggests that the regulators should instead help consumers learn whether their data have been collected by forcing the firm to reveal its data collection policy rather than the results of its data analysis. This implication relies, of course, on our assumption that consumers are perfectly rational, aware of firm's data collection, and able to make inferences from firm choices of product menus and pricing schemes.

To be sure, much of the debate on data collection by commercial interests revolves around the issue of consumers' naïveté. If data are collected without the consumers' explicit consent or knowledge, then the pure strategy equilibrium from the main model may not hold; instead, a mixed strategy may be more appropriate when consumers need to infer whether data collection has occurred from the observed product line. It is also clear from previous research that if consumers have perfect knowledge of their preferences, then consumer data collection can reduce their surplus (e.g., Fudenberg and Villas-Boas 2006). Therefore, our findings that consumers can benefit from sharing their data with for-profit firms are shown under the specific conditions of uninformed consumer preference.

We acknowledge that our model has overlooked some important issues regarding data collection and consumers' concern about their privacy. We have focused on the firm's collection of anonymous consumer data for product design and pricing. But consumers' concerns for privacy extend beyond protecting their information rents. Consumers are also suspicious that the data are not truly anonymous and could be acquired by third parties with ill intent. With these caveats in mind, our hope with this research is to show that there is still much research needed to understand the full implications of consumer data collection.

In looking toward future research, it is worth mentioning that the notion of superior knowledge extends beyond the product line question studied here. For example, when the firm uses personalized discounts, the catalog, or list, price may play an important role in communicating to consumers who have uncertain preferences.[14] In this case, the key aspect of superior knowledge also regards consumers who are a priori unsure in which preference segment they reside. Therefore, studies of superior knowledge may have broader implications than product line design whenever consumers are subject to perceptual errors in assessing their preferences.

## Acknowledgments

The authors thank senior editor, Yuxin Chen; associate editor, Juanjuan Zhang; and two anonymous reviewers for their excellent suggestions. They also gratefully acknowledge helpful comments from Matt Selove, Guofu Tan, Dina Mayzlin, Shantanu Dutta, Bill Zame; seminar participants at University of Texas at Dallas's Jindal School, UCLA Department of Economics, University of Hong Kong, Peking University, Fudan University, Xiamen University, Shanghai University of Finance and Economics; and attendees of the 2018 Marketing Science Conference. The authors also thank the National Natural Science Foundation of China for financial support.

## Appendix. Proofs of Propositions

### Proof of Lemma 1

Consider the case that both types of consumers observe the same signal under some market states and are informed under others. Let the probability that both types observe the same signal be $\beta$. Denote the firm's profit when it does not collect data as $\Pi^N$ and when it collects data as $\Pi^C$.

When the firm does not collect data, there are three segments of consumers in expectation: informed $H$-types, uninformed consumers, and informed $L$-types. Suppose that the parameter space allows full market coverage; then the firm's optimization problem is thus as follows:

$$\max_{\{p_1,p_2,p_3,q_1,q_2,q_3\}} \beta\lambda\left(p_1-\frac{q_1^2}{2}\right)+(1-\beta)\left(p_2-\frac{q_2^2}{2}\right)+\beta(1-\lambda)\left(p_3-\frac{q_3^2}{2}\right)$$

subject to $\alpha_j q_j - p_j \geq \max\{\alpha_j q_k - p_k, 0\}$ for every $k, j = 1, 2, 3$, and $k \neq j$, where $\alpha_1 = \alpha_H$; $\alpha_2 = e \equiv \lambda\alpha_H + (1-\lambda)\alpha_L$; $\alpha_3 = \alpha_L$, and $p_j, q_j \geq 0$. Denote the equilibrium profit and price as $p_j^*, q_j^*$. Then because $e q_2^* - p_2^* \geq e q_3^* - p_3^* = (e - \alpha_L) q_3^* + (\alpha_L q_3^* - p_3^*) \geq (e - \alpha_L) q_3^* > 0$, we must have

$$\Pi^N < \Pi(p_1 = p_1^*, q_1 = q_1^*, p_2 = e\, q_2^*, q_2 = q_2^*, p_3 = p_3^*, q_3 = q_3^*).$$

By contrast, when a firm collects data, it can design two product lines in each situation when the consumers are informed and uninformed. The firm's optimization problem is as follows:

$$\max_{\{p_1, p_2, p_3, q_1, q_2, q_3\}} \beta\lambda\left(p_1 - \frac{q_1^2}{2}\right) + (1-\beta)\left(p_2 - \frac{q_2^2}{2}\right) + \beta(1-\lambda)\left(p_3 - \frac{q_3^2}{2}\right)$$

subject to $\alpha_j q_j - p_j \geq \max\{\alpha_j q_k - p_k, 0\}$ for every $k, j = 1, 3$, and $k \neq j$, where $\alpha_1 = \alpha_H$; $\alpha_3 = \alpha_L$; $e\, q_2 - p_2 \geq 0$; and $p_j, q_j \geq 0$. Denote the equilibrium profit and price as $p_j^{**}, q_j^{**}$. Because the rational participation constraint is binding for product 2 when both consumers are uninformed, the firm's equilibrium profit is

$$\Pi^C = \Pi(p_1^{**}, q_1^{**}, p_2 = e\, q_2^{**}, q_2^{**}, p_3^{**}, q_3^{**}).$$

Because the restrictions for $p_j^*, q_j^*$ are stronger than those for $p_j^{**}, q_j^{**}$, we must have $\Pi(p_1^*, q_1^*, p_2 = e\, q_2^*, q_2^*, p_3^*, q_3^*) < \Pi(p_1^{**}, q_1^{**}, p_2 = e\, q_2^{**}, q_2^{**}, p_3^{**}, q_3^{**})$; therefore, $\Pi^N < \Pi^C$. The firm is always better off from data collection when it does not create superior knowledge.

It remains to show that data collection is not a Pareto improvement. It suffices to show that $L$-types are strictly worse off from data collection. First consider the case when the firm does not collect data. The $L$-types' expected surplus is $CS^N = (1-\beta)(\alpha_L q_2^* - p_2^*) + \beta(\alpha_L q_3^* - p_3^*)$. The second term is 0 because the rationality participation constraint is binding for the informed $L$-types. The first term equals $(1-\beta) \cdot [(e q_2^* - p_2^*) - (e - \alpha_L) q_2^*] > (\beta - 1)(e - \alpha_L) q_2^*$, because $e q_2^* - p_2^* > 0$. Because the incentive compatibility constraint for the informed $H$-types requires that the quality of product 2 to be downwardly distorted, we must have $q_2^* < e$; thus $CS^N > (\beta - 1)(e - \alpha_L) e$.

Second, consider the case when the firm collects data. The $L$-types' expected surplus is $CS^C = (1-\beta)(\alpha_L q_2^{**} - p_2^{**}) + \beta(\alpha_L q_3^{**} - p_3^{**})$. Similarly, the second term is 0. Because the rational participation constraint is binding for product 2, we have $q_2^{**} = e$ and $p_2^{**} = e^2$, so $CS^C = (\beta - 1)(e - \alpha_L) e$. Therefore, we must have $CS^N > CS^C$. That is, $L$-types are strictly worse off with data collection. ■

## Proof of Lemma 2

We prove by contradiction. Suppose that for any $m \in (0, 1)$ there does not exist another $m' \in (0, 1)$ such that $\theta_H(m) = \theta_L(m')$. Then, by Assumption 2, because $\theta_L(.)$ is continuous, we must have either (1) $\theta_H(m) > \theta_L(m')$ for all $m' \in (0, 1)$ or (2) $\theta_H(m) < \theta_L(m')$ for all $m' \in (0, 1)$. The second case implies that $\theta_H(m) < \theta_L(m)$, which is contradictory to Assumption 3. Because $\theta_H(0) < \theta_L(1)$, and $\theta_i(.)$ are continuous, then for an arbitrarily small $0 < \epsilon < \frac{\theta_L(1) - \theta_H(0)}{2}$, there must exist $m, m' \in (0, 1)$ such that $\theta_H(m) < \theta_H(0) + \epsilon$ and $\theta_L(m') > \theta_L(1) - \epsilon$. So $\theta_L(m') > \frac{\theta_L(1) + \theta_H(0)}{2} > \theta_H(m)$, which contradicts with the first case. Therefore, there must exist $m, m' \in (0, 1)$ such that $\theta_H(m) = \theta_L(m')$.

Because, by Assumption 2, $\theta_L(m') < \theta_L(1)$, we have $\theta_H(m) < \theta_L(1)$. Because $\theta$ is increasing, this condition is equivalent to $m < \theta_H^{-1} \circ \theta_L(1)$, so $H$-types are uninformed under $m \in (0, \theta_H^{-1} \circ \theta_L(1))$. Similarly, because $\theta_H(m) > \theta_H(0)$ by Assumption 2, we have $\theta_L(m') > \theta_H(0)$, thus $L$-types are uninformed if and only if $m \in (\theta_L^{-1} \circ \theta_H(0), 1)$. Therefore, those intervals are where data aggregation creates superior knowledge. ■

To prove the propositions in the main text, we first establish that the mapping between any uninformed consumers' inferred type $(m_H, m_L)$ and the segments $(\mathcal{M}^1, \mathcal{M}^0)$ in Lemma B.1 is bijective, given the stylized restriction of basic model that $T(1) \leq T^{-1}(0)$, where $T(m) \equiv \theta_H^{-1} \circ \theta_L(m)$, $\mathcal{M}^0 = [T(1), 1)$, and $\mathcal{M}^1 = (0, T(1))$.

**Lemma B.1.** *Suppose that $T(1) \leq T^{-1}(0)$. Any uninformed consumer infers the market state $m$ as either $m_H \in \mathcal{M}^1$, under which she is an H-type, or $m_L \in \mathcal{M}^0$, under which she is an L-type.*

**Proof.** Because $m_H = T(m_L)$ and $m_L < 1$, by Lemma A1 in the online appendix, we must have $m_H < T(1)$; thus $m_H \in \mathcal{M}^1$. Similarly, because $m_L = T^{-1}(m_H)$ and $m_H > 0$, we must have $m_L > T^{-1}(0)$. Because the basic model assumes $T(1) \leq T^{-1}(0)$, we must have $m_L \geq T(1)$; thus $m_L \in \mathcal{M}^0$. ■

Lemma B.1 helps simplify the proof to the following propositions.

## Proof of Proposition 1

First show that the given equilibrium satisfies sequential rationality and Bayesian rules when $m \in \mathcal{M}^0$ (Step 1). Then show the same for $m \in \mathcal{M}^1$, and the equilibrium is Pareto efficient (Step 2). Finally, we show that the equilibrium uniquely survives both the intuitive criterion and D1 (Step 3).

**Step 1.** When $m \in \mathcal{M}^0$, only $L$-types are uninformed. Consider an arbitrary deviation $\Psi' \neq \Psi^*$. Because $\Psi^*$ is firm optimal when $L$-types learn their type, $\Psi'$ is profitable only if $L$-types overestimate their type (i.e., $\mu_i(\Psi') > 0$). By (iii), $\mu_i(\Psi') > 0$ only if $\hat{\Pi}(\mathcal{M}^0, r_i, \Psi') \leq \Pi_F^*$, where $r_i \in BR(\Psi', \mu_i = 1)$. Because $\mu_i < 1$ indicates that the consumer attributes positive probability that $m \in \mathcal{M}^0$ or, equivalently, that she is an $L$-type, $\mu_i = 1$ is the most favorable belief for the firm (i.e., $\hat{\Pi}(\mathcal{M}^0, r_i, \Psi') \geq \hat{\Pi}(\mathcal{M}^0, r_i', \Psi')$ for any $r_i' \in BR(\Psi', \mu_i \leq 1)$). Therefore, $\hat{\Pi}(\mathcal{M}^0, r_i', \Psi') \leq \Pi_F^*$: no deviation $\Psi'$ is more profitable than $\Psi^*$ for $\mu_i(\Psi') > 0$. In addition, $r_L^* = 1$ is a best response for $L$-types, because $p_2^*(\mathcal{M}^0) = \alpha_L q_2$ whenever it is available. Therefore, the equilibrium satisfies the sequential rationality conditions. Finally, because $\Psi^*$ is offered only when $m \in \mathcal{M}^0$, the updated belief $\mu_i = 0$ satisfies Bayes' rule.

**Step 2.** When $m \in \mathcal{M}^1$, only $H$-types are uninformed. To infer the firm's private information of the market state, $H$-types need to estimate $\hat{\Pi}(\mathcal{M}^0, r_i, \Psi')$. Because the uninformed consumers would assume that the other consumers under $\mathcal{M}^0$ would be the informed $H$-types and thus choose the high-end product, provided that the product line satisfies the incentive compatibility constraint, the estimated firm's deviation profit should be $\hat{\Pi}(\mathcal{M}^0, r_i, \Psi') = p_1' - \frac{(q_1')^2}{2}$, where $r_i \in BR(\Psi',$

$\mu_i = 1$). By (iii) $\mu_i = 1$ if and only if $p_1' - \frac{(q_1')^2}{2} \le \Pi_F^*$. Therefore, this additional restriction poses a signaling constraint, in addition to the rational participation and incentive compatibility constraints in the standard product line design.

**Step 2.1.** Suppose that $\lambda < \alpha_L/\alpha_H$; then the firm's profit in the full-information standard model is $\Pi_F^* = \frac{\lambda(\alpha_H-\alpha_L)^2}{2(1-\lambda)} + \frac{\alpha_L^2}{2}$. By (iii), the firm's maximization problem is

$$\max_{\{q_1,q_2,p_1,p_2\}} \lambda\left(p_1 - \frac{(q_1)^2}{2}\right) + (1-\lambda)\left(p_2 - \frac{(q_2)^2}{2}\right)$$

subject to

(A.1a): $\alpha_H q_1 - p_1 \ge \alpha_H q_2 - p_2$;
(A.1b): $\alpha_H q_1 - p_1 \ge 0$;
(A.2a): $\alpha_L q_2 - p_2 \ge \alpha_L q_1 - p_1$;
(A.2b): $\alpha_L q_2 - p_2 \ge 0$;
(A.3): $p_1 - \frac{(q_1)^2}{2} \le \frac{\lambda(\alpha_H-\alpha_L)^2}{2(1-\lambda)} + \frac{\alpha_L^2}{2}$.

Notice from the inequality (A.3) that the nonlinear constrained optimization problem can be solved by convex programming. First, we show (A.3) is binding at the optimized solution. We prove by contradiction: suppose that the constraint (A.3) is slack at the optimized solution (i.e., $p_1^* - \frac{(q_1^*)^2}{2} < \frac{\lambda(\alpha_H-\alpha_L)^2}{2(1-\lambda)} + \frac{\alpha_L^2}{2}$); then the optimization problem is equivalent to the full-information case, and the optimal product line is thus $q_1^* = \alpha_H$; $p_1^* = \frac{(\alpha_H-\alpha_L)^2}{1-\lambda} + \alpha_H\alpha_L$. However, $p_1^* - \frac{(q_1^*)^2}{2} = \left(\frac{\lambda(\alpha_H-\alpha_L)^2}{2(1-\lambda)} + \frac{\alpha_L^2}{2}\right) + \frac{(\alpha_H-\alpha_L)^2}{2(1-\lambda)} > \Pi_F^*$, which contradicts with the assumption that the constraint (A.3) is slack at the optimized solution. Therefore, (A.3) must be binding at the optimized solution.

Second, we show that (A.2a) cannot be binding, because it is suboptimal for the firm to bind (A.2a). To see this, note that (A.2a) is the only floor for $p_1$. Suppose $p_1^* = p_2^* + \alpha_L(q_1^* - q_2^*)$, and the other inequalities hold. Then the firm is always better off by setting $p_1 = p_1^* + \varepsilon$, where $\varepsilon > 0$ is sufficiently small such that the other inequalities still hold.

Third, we show that (A.2b) must be binding. Because (A.2b) is equivalent to $p_2 \le \alpha_L q_2$, which is a price ceiling for the low-end product, but because (A.2a) is not binding, (A.2b) is the only ceiling for $p_2$. Because it is optimal for the firm to charge $p_2$ as high as possible, the price ceiling must be binding.

From the above discussion, the optimization problem can be simplified using $p_2 = \alpha_L q_2$ and $p_1 = \frac{\lambda(\alpha_H-\alpha_L)^2}{2(1-\lambda)} + \frac{\alpha_L^2}{2} + \frac{(q_1)^2}{2}$. Therefore, (A.1a) is equivalent to $q_2 \le \frac{\alpha_H(1-2\lambda)+\alpha_L}{2(1-\lambda)} - \frac{(\alpha_H-q_1)^2}{2(\alpha_H-\alpha_L)}$. The above problem can thus be simplified as follows:

$$\max_{\{q_1,q_2\}} \lambda\Pi_F^* + (1-\lambda)\left(\alpha_L q_2 - \frac{(q_2)^2}{2}\right)$$

subject to $q_2 \le \frac{\alpha_H(1-2\lambda)+\alpha_L}{2(1-\lambda)} - \frac{(\alpha_H-q_1)^2}{2(\alpha_H-\alpha_L)}$.

Suppose that the first-order conditions, $q_2^* = \alpha_L$, satisfy the constraint that $q_2 \le \frac{\alpha_H(1-2\lambda)+\alpha_L}{2(1-\lambda)} - \frac{(\alpha_H-q_1)^2}{2(\alpha_H-\alpha_L)}$. Substitute $q_2^* = \alpha_L$ into the constraint to obtain $1 - 2\lambda \ge \frac{(1-\lambda)(\alpha_H-q_1)^2}{(\alpha_H-\alpha_L)^2}$. We discuss whether the previous condition holds in the following two scenarios.

If $\lambda \le \frac{1}{2}$, then $1 - 2\lambda \ge 0 = \frac{(1-\lambda)(\alpha_H-q_1)^2}{(\alpha_H-\alpha_L)^2}$ when setting $q_1^* = \alpha_H$. In this case, the equilibrium product line is $q_2^* = \alpha_L$, $p_2^* = \alpha_L^2$; $q_1^* = \alpha_H$, $p_1^* = \frac{1}{2(1-\lambda)}(\alpha_H^2 - 2\lambda\alpha_H\alpha_L + \alpha_L^2)$.

If $\lambda > \frac{1}{2}$, then because $1 - 2\lambda < 0 \le \frac{(1-\lambda)(\alpha_H-q_1)^2}{(\alpha_H-\alpha_L)^2}$, $q_2^* = \alpha_L$ cannot be a feasible solution. In this case, both constraints (A.1a) and (A.3) must be binding. The intuition is that the constraint (A.3) imposes a price ceiling on $p_1^*$, and (A.1a) imposes a quality ceiling on $q_2$. Therefore, unless the price ceiling is low enough, the quality ceiling must be lower than the efficient level.

When both (A.1a) and (A.3) are binding, we reduce the maximand by substituting the binding condition of $q_2$ into the profit maximization:

$$\max_{\{q_1,q_2\}} \lambda\Pi_F^* + (1-\lambda)\left(\alpha_L q_2 - \frac{(q_2)^2}{2}\right)$$

$$\text{subject to } q_2 = \frac{\alpha_H(1-2\lambda)+\alpha_L}{2(1-\lambda)} - \frac{(\alpha_H-q_1)^2}{2(\alpha_H-\alpha_L)}.$$

We solve the above maximization problem using the Lagrangian method as $q_1^* = \alpha_H$ and $q_2^* = \frac{\alpha_H(1-2\lambda)+\alpha_L}{2(1-\lambda)}$, where $q_2^* > 0$, because $\alpha_H\lambda < \alpha_L$ implies $\alpha_H(1-2\lambda) + \alpha_L > \alpha_H(1-\lambda) > 0$.

**Step 2.2.** Suppose that $\lambda \ge \alpha$; then $\Pi_F^* = \lambda\frac{\alpha_H^2}{2}$. The firm's optimization problem is

$$\max_{\{q_1,q_2,p_1,p_2\}} \lambda\left(p_1 - \frac{(q_1)^2}{2}\right) + (1-\lambda)\left(p_2 - \frac{(q_2)^2}{2}\right)$$

subject to (A.1a)–(A.2b) and

(A.4): $p_1 - ((q_1)^2/2) \le \lambda(\alpha_H^2)/2$

Similarly, constraint (A.4) needs to be binding. Otherwise, we have $q_1^* = \alpha_H$, $p_1^* = \alpha_H^2$ and thus $p_1^* - \frac{(q_1^*)^2}{2} > \lambda\frac{\alpha_H^2}{2}$, contradicting with constraint (A.4). Moreover, we show that if $\lambda < \frac{1}{2}$, constraint (A.1a) is not binding. Therefore, we have the first-best solution: $q_2^* = \alpha_L$; $q_1^* = \alpha_H$; $p_2^* = \alpha_L^2$; and $p_1^* = \frac{(1+\lambda)}{2}\alpha_H^2$.

If $\lambda \ge \frac{1}{2}$, then the cannibalization constraint (A.1a) must be binding. Therefore, we can simplify the profit maximization problem as follows:

$$\max_{\{q_H,q_L\}} \lambda\left(\lambda\frac{\alpha_H^2}{2}\right) + (1-\lambda)\left(\alpha_L q_2 - \frac{(q_2)^2}{2}\right)$$

subject to $\lambda\frac{\alpha_H^2}{2} + \frac{(q_1)^2}{2} - \alpha_L q_2 - \alpha_H(q_1 - q_2) = 0$.

Use again the Lagrangian method to obtain $q_1^* = \alpha_H$; $p_1^* = \frac{(1+\lambda)}{2}\alpha_H^2$; $q_2^* = \frac{(1-\lambda)\alpha_H^2}{2(\alpha_H-\alpha_L)}$; $p_2^* = \frac{(1-\lambda)\alpha_H^2\alpha_L}{2}$.

**Step 2.3.** Because the equilibrium product line in (i) satisfies the participation constraint and incentive compatibility constraint, $r_H^*$ defined in (ii) is a best response. Therefore, the equilibrium satisfies a consumer's sequential rationality condition. Furthermore, by (i), the product line is separating on the market state; therefore the updated belief $\mu_i$ is either 1 or 0 and satisfies Bayesian rule.

**Step 3.** We need to show that the equilibrium uniquely survives both intuitive criterion and D1. Because D1 is stronger

than IC, it suffices to show that the equilibrium satisfies D1 and is unique with IC. We show the first part in Step 3.1 and the second part in Step 3.2.

**Step 3.1.** By the definition of D1 (Banks and Sobel 1987, Cho and Kreps 1987, Cho and Sobel 1990), it is equivalent to show that for any out-of-equilibrium pricing strategy $\Psi' \neq \Psi^*$ such that whenever the following condition holds for any $i \in \{H, L\}$ and $j \neq k \in \{\mathcal{M}^0, \mathcal{M}^1\}$:

$$\cup_\mu \{r_i | \Pi^*(j) \leq \hat{\Pi}_i(j, \Psi', r_i)\} \subsetneq \cup_\mu \{r_i | \Pi^*(k) < \hat{\Pi}_i(k, \Psi', r_i)\},$$

where $r_i$ is a best response to $\Psi'$ given an arbitrary belief $\mu \in [0,1]$ [i.e., $r_i = BR(\Psi', \mu \in [0,1])$], consumer $i$ updates her belief as follows:

$$\mu_i(\Psi') = \begin{cases} 1 & \text{if } j \in \mathcal{M}^0, \\ 0 & \text{if } j \in \mathcal{M}^1. \end{cases}$$

Suppose that $j \in \mathcal{M}^0$ and $k \in \mathcal{M}^1$; then for any $\Psi'$ whenever the above condition holds, the set $\cup_\mu \{r_i | \Pi^*(\mathcal{M}^1) < \hat{\Pi}_i(\mathcal{M}^1, \Psi', r_i)\} \neq \emptyset$, which then implies $\mu(\Psi') = 1$ by the belief rule.

Next, suppose that $j \in \mathcal{M}^1$ and $k \in \mathcal{M}^0$. We want to show that $\mu(\Psi') = 0$ whenever $\cup_\mu \{r_i | \Pi^*(\mathcal{M}^1) \leq \hat{\Pi}_i(\mathcal{M}^1, \Psi', r_i)\} \subsetneq \cup_\mu \{r_i | \Pi^*(\mathcal{M}^0) < \hat{\Pi}_i(\mathcal{M}^0, \Psi', r_i)\}$. We prove by contradiction: if $\mu(\Psi') \neq 0$, then by part (iii) of Proposition 1, we must have $\mu(\Psi') = 1$; thus $\Pi^*(\mathcal{M}^1) \leq \hat{\Pi}_i(\mathcal{M}^1, \Psi', r_i)$. Because the complementary set is null (i.e., $\cup_\mu \{r_i | \Pi^*(\mathcal{M}^1) > \hat{\Pi}_i(\mathcal{M}^1, \Psi', r_i)\} = \emptyset$), it violates the strict inclusion condition that $\cup_\mu \{r_i | \Pi^*(\mathcal{M}^1) \leq \hat{\Pi}_i(\mathcal{M}^1, \Psi', r_i)\} \subsetneq \cup_\mu \{r_i | \Pi^*(\mathcal{M}^0) < \hat{\Pi}_i(\mathcal{M}^0, \Psi', r_i)\}$. Therefore, $\mu(\Psi') = 0$ whenever the condition holds for $j \in \mathcal{M}^1$. Thus, the equilibrium survives D1.

**Step 3.2.** Given the property of the market space, there does not exist another separating equilibrium that survives IC. Therefore, it remains to show that any pooling equilibrium fails the intuitive criterion. Denote $(\tilde{\Psi}, \tilde{r}_i, \tilde{\mu})_{i=H,L}$ as a pooling equilibrium such that $\tilde{\Psi}(\mathcal{M}^0) = \tilde{\Psi}(\mathcal{M}^1)$ and $\tilde{\mu}_H(\tilde{\Psi}) = \tilde{\mu}_L(\tilde{\Psi}) > 0$.

We consider two types of pooling product lines. The first type has two products: $\tilde{\Psi} = \{\tilde{q}_1, \tilde{p}_1, \tilde{q}_2, \tilde{p}_2\}$, such that (1) both the informed $H$-types and uninformed $L$-types purchase the $H$-product under $\mathcal{M}^0$, and (2) the unformed $H$-types purchase the $H$-product and informed $L$-types purchase the $L$-product under $\mathcal{M}^1$. Then the equilibrium profit in each market state is $\tilde{\Pi}(\mathcal{M}^0) = \tilde{p}_1 - \frac{\tilde{q}_1^2}{2}$, and $\tilde{\Pi}(\mathcal{M}^1) = \lambda\left(\tilde{p}_1 - \frac{\tilde{q}_1^2}{2}\right) + (1-\lambda)\left(\tilde{p}_2 - \frac{\tilde{q}_2^2}{2}\right)$. Denote the uninformed consumers' estimate as $e \equiv E[\alpha|\theta]$. The optimal pooling product line is designed as $\tilde{q}_1 = e$, $\tilde{p}_1 = \frac{(e-\alpha_L)^2}{1-\lambda} + e\alpha_L$, $\tilde{q}_2 = \frac{\alpha_L - \lambda e}{1-\lambda}$, $\tilde{p}_2 = \frac{\alpha_L^2 - \lambda e \alpha_L}{1-\lambda}$.

Denote $\Psi'$ as an out-of-equilibrium product line (i.e., $\Psi' = \{q_1', p_1', q_2', p_2'\}$), where $p_1'$ satisfies $p_1' - \frac{(q_1')^2}{2} = \tilde{p}_1 - \frac{\tilde{q}_1^2}{2}$. Therefore, the uninformed consumers know that the incentives of the deviation are absent under $\mathcal{M}^0$. It remains to show that the deviation is profitable under $\mathcal{M}^1$, because in this case, uninformed consumers with an IC belief will update estimates by assigning probability 1 to $\mathcal{M}^1$ and thus purchase the $H$-product in $\Psi'$, so we can reject that $\tilde{\Psi}$ is a pooling strategy by IC. Let $q_1' = \alpha_H$. Because the rational participation constraint is binding, we have $\alpha_L q_2' = p_2'$. By the incentive compatibility constraint, $\alpha_H q_2' - p_2' \leq \alpha_H q_1' - p_1'$. Because $p_1' - \frac{\alpha_H^2}{2} = \tilde{p}_1 - \frac{\tilde{q}_1^2}{2} = \frac{(e-\alpha_L)^2}{1-\lambda} + e\alpha_L - \frac{e^2}{2}$, we obtain that $q_2' \leq \frac{\alpha_L + \alpha_H}{2} + \frac{(e-\alpha_L)^2}{2(\alpha_H - \alpha_L)} \frac{1+\lambda}{1-\lambda}$, which is greater than $\alpha_L$. Therefore, it is always possible to set $q_2' = \alpha_L$ to satisfy the above conditions. In this case, the firm's profit under $\mathcal{M}^1$ is $\Pi'(\mathcal{M}^1) = \lambda\left(p_1' - \frac{\alpha_H^2}{2}\right) + (1-\lambda)\frac{\alpha_L^2}{2}$, so $\Pi'(\mathcal{M}^1) > \tilde{\Pi}(\mathcal{M}^1)$ is equivalent to $\frac{(\alpha_L - \tilde{q}_2)^2}{2} > 0$. Therefore, $\Psi'$ is profitable only under $\mathcal{M}^1$. Consequently, $\tilde{\Psi}$ as a pooling strategy fails the intuitive criterion.

The second type we examine has three products: $\tilde{\Psi} = \{\tilde{q}_1, \tilde{p}_1, \tilde{q}_2, \tilde{p}_2, \tilde{q}_3, \tilde{p}_3\}$, where $\tilde{q}_1 \geq \tilde{q}_2 \geq \tilde{q}_3$, such that the informed $H$-types purchase $\tilde{q}_1$, uninformed consumers purchase $\tilde{q}_2$, and the informed $L$-types purchase $\tilde{q}_3$. Then the equilibrium profits are $\tilde{\Pi}(\mathcal{M}^0) = \lambda\left(\tilde{p}_1 - \frac{\tilde{q}_1^2}{2}\right) + (1-\lambda)\left(\tilde{p}_2 - \frac{\tilde{q}_2^2}{2}\right)$, and $\tilde{\Pi}(\mathcal{M}^1) = \lambda\left(\tilde{p}_2 - \frac{\tilde{q}_2^2}{2}\right) + (1-\lambda)\left(\tilde{p}_3 - \frac{\tilde{q}_3^2}{2}\right)$. Again, denote $\Psi'$ as an out-of-equilibrium product line, where $p_1'$ satisfies $p_1' - \frac{\alpha_H^2}{2} = \tilde{\Pi}(\mathcal{M}^0)$. Similar to the above analysis, the deviation has no incentive under $\mathcal{M}^0$. It remains to show that $\Pi'(\mathcal{M}^1) > \tilde{\Pi}(\mathcal{M}^1)$. Because $\tilde{p}_1 - \frac{\tilde{q}_1^2}{2} > \tilde{p}_2 - \frac{\tilde{q}_2^2}{2} > \tilde{p}_3 - \frac{\tilde{q}_3^2}{2}$, we have $\Pi'(\mathcal{M}^1) = \lambda\left(p_1' - \frac{\alpha_H^2}{2}\right) + (1-\lambda) \cdot \left(p_3' - \frac{q_3'^2}{2}\right) > \lambda\left(\tilde{p}_2 - \frac{\tilde{q}_2^2}{2}\right) + (1-\lambda)\left(p_3' - \frac{q_3'^2}{2}\right)$. But because the incentive compatibility constraint is stronger with the pooling product line, the margin for the lowest-quality product much be smaller than that in the deviating product line; thus $p_3' - \frac{q_3'^2}{2} > \tilde{p}_3 - \frac{\tilde{q}_3^2}{2}$. Therefore, $\Pi'(\mathcal{M}^1) > \tilde{\Pi}(\mathcal{M}^1)$.

In summary, both types of pooling equilibrium fail IC. The separating equilibrium in Proposition 3 uniquely survives IC. Because D1 is stronger, the equilibrium also uniquely survives D1. ■

## Proof of Proposition 2

Because $q_2^*(\mathcal{M}^0)$ is the same as the equilibrium under full information, it suffices to check whether $q_2^*(\mathcal{M}^1) > q_2^*(\mathcal{M}^0)$ to show that consumers obtain higher-quality products (on average).

(1) If $\lambda \leq \frac{1}{2}$, then $q_2^*(\mathcal{M}^1) = \alpha_L > q_2^*(\mathcal{M}^0)$.

(2) If $\alpha > \lambda$, and $\lambda > \frac{1}{2}$, then $q_2^*(\mathcal{M}^1) = \frac{\alpha_H(1-2\lambda) + \alpha_L}{2(1-\lambda)} = \frac{\alpha_H - \alpha_L}{2(1-\lambda)} + q_2^*(\mathcal{M}^0) > q_2^*(\mathcal{M}^0)$.

(3) If $\alpha \leq \lambda$, and $\lambda > \frac{1}{2}$, then $q_2^*(\mathcal{M}^1) = \frac{(1-\lambda)\alpha_H^2}{2(\alpha_H - \alpha_L)} = \frac{\alpha_H^2[(1-\alpha)^2 + (\alpha-\lambda)^2]}{2(1-\lambda)(\alpha_H - \alpha_L)} + q_2^*(\mathcal{M}^0) > q_2^*(\mathcal{M}^0)$.

It remains to show that $q_2^*(\mathcal{M}^1) \leq \alpha_L$. First, $\frac{\alpha_H(1-2\lambda) + \alpha_L}{2(1-\lambda)} - \alpha_L = \frac{\alpha_H - \alpha_L}{2(1-\lambda)}(1-2\lambda)$. Because $q_2^*(\mathcal{M}^1) = \frac{\alpha_H(1-2\lambda) + \alpha_L}{2(1-\lambda)}$ is feasible when $\lambda \geq \frac{1}{2}$, we have $q_2^*(\mathcal{M}^1) \leq \alpha_L$. Second, $\frac{(1-\lambda)\alpha_H^2}{2(\alpha_H - \alpha_L)} - \alpha_L = \frac{\alpha_H^2}{2(\alpha_H - \alpha_L)} \cdot [(1-2\lambda)(1-\lambda) + (\alpha - \lambda)]$. Because $q_2^*(\mathcal{M}^1) = \frac{(1-\lambda)\alpha_H^2}{2(\alpha_H - \alpha_L)}$ is feasible when $\alpha \leq \lambda$, and $\lambda > \frac{1}{2}$, we have $q_2^*(\mathcal{M}^1) < \alpha_L$. ■

## Proof of Proposition 3

Solve the optimization problem in Equation (1). Note that if $q_3 > 0$, then the constraints are equivalent to $\alpha_H q_1 - p_1 = \alpha_H q_2 - p_2$, $e q_2 - p_2 = e q_3 - p_3$, and $e q_3 = p_3$. Substitute the price-conditional functions to the objective function and derive the first-order conditions on qualities to obtain that $\hat{q}_1 = \alpha_H$, $\hat{q}_2 = \alpha_L + \lambda^2(\alpha_H - \alpha_L)$, and $\hat{q}_3 = \alpha_L - \frac{\lambda(1+\lambda)}{1-\lambda}(\alpha_H - \alpha_L)$. Clearly, $\hat{q}_1 > 0$ and $\hat{q}_2 > 0$. It remains to derive the firm's product line strategy in the parameter space where $\hat{q}_3 \leq 0$, or equivalently, $\alpha \leq \frac{\lambda + \lambda^2}{1 + \lambda^2}$.

If $\alpha \leq \frac{\lambda + \lambda^2}{1 + \lambda^2}$, then because the profit function can be rewritten as a quadratic function of $\hat{q}_3$ (given the other parameters

satisfying the restrictions), the corner solution must be that $\hat{q}_3 = 0$. Therefore, the firm targets only informed $H$-types and the uninformed consumers. The profit maximization problem becomes

$$\max_{\{p_1,p_2,q_1,q_2\}} \frac{\lambda}{2}\left(p_1 - \frac{q_1^2}{2}\right) + \frac{1}{2}\left(p_2 - \frac{q_2^2}{2}\right)$$

subject to (1) $\alpha_H q_1 - p_1 \geq \max\{\alpha_H q_2 - p_2, 0\}$,
(2) $e q_2 - p_2 \geq \max\{e q_1 - p_1, 0\}$,
(3) $p_1, p_2, q_1, q_2 \geq 0$.

Solving for the maximizer, $\hat{q}_1 = \alpha_H$, $\hat{q}_2 = \alpha_L + \lambda^2(\alpha_H - \alpha_L)$, $\hat{p}_1 = \alpha_H(\hat{q}_1 - \hat{q}_2) + \hat{p}_2$, and $\hat{p}_2 = e\hat{q}_2$. ■

## Proof of Proposition 4

For the incentives of data collection, it suffices to check $(\Pi_c^* - \Pi_{nc}^*)$ in the following six regions:

(1) If $\alpha > \frac{\lambda+\lambda^2}{1+\lambda^2}$ and $\lambda \leq \frac{1}{2}$, then
$\Pi_c^* - \Pi_{nc}^* = \frac{(1-\alpha)^2\lambda^2}{4(1-\lambda)}A$, where $A \equiv 1 - 2\lambda - 2\lambda^2 + \lambda^3$.

(2) If $\alpha > \frac{\lambda+\lambda^2}{1+\lambda^2}$ and $\lambda > \frac{1}{2}$, then
$\Pi_c^* - \Pi_{nc}^* = \frac{(1-\alpha)^2}{16(1-\lambda)}[4\lambda^2 A - (2\lambda-1)^2]$.

(3) If $\alpha \in (\lambda, \frac{\lambda+\lambda^2}{1+\lambda^2}]$ and $\lambda \leq \frac{1}{2}$, then
$\Pi_c^* - \Pi_{nc}^* = \frac{1}{4(1-\lambda)}[\lambda^2 A(1-\alpha)^2 + ((\lambda^2+1)\alpha - (\lambda^2+\lambda))^2]$.

(4) If $\alpha \in (\lambda, \frac{\lambda+\lambda^2}{1+\lambda^2}]$ and $\lambda > \frac{1}{2}$, then
$\Pi_c^* - \Pi_{nc}^* = \frac{1}{16(1-\lambda)B}[((\alpha-1)B + 4(1-\lambda)(1+\lambda^2))^2 + 4(1-\lambda)^2(4\lambda^2 A - (2\lambda-1)^2)]$
where $B \equiv 3 + 4\lambda + 8\lambda^2 - 8\lambda^3 - 4\lambda^4 + 4\lambda^5 > 0$.

(5) If $\alpha \leq \lambda \leq \frac{1}{2}$, then
$\Pi_c^* - \Pi_{nc}^* = \frac{\lambda(1-\lambda)}{4}[\lambda(1+\lambda)(1-\alpha)^2 - \alpha^2]$.

(6) If $\alpha \leq \lambda$ and $\lambda > \frac{1}{2}$, then
$\Pi_c^* - \Pi_{nc}^* = \frac{1}{16(1-\lambda)}[-4\alpha^4(-1+\lambda^2)^2 + 8\alpha^3(1-3\lambda^2+2\lambda^4) - 8\alpha^2(1-\lambda-3\lambda^2+3\lambda^4) + 4\alpha(1-2\lambda-3\lambda^2+4\lambda^4) - 1 + 3\lambda + \lambda^2 + \lambda^3 - 4\lambda^4]$.

Because $\lambda \in (0,1)$, $A \equiv 1 - 2\lambda - 2\lambda^2 + \lambda^3 > 0$, if and only if $\lambda < \frac{3-\sqrt{5}}{2} \approx 0.38$. Thus, in item (1), $\Pi_c^* > \Pi_{nc}^*$ is equivalent to $\lambda < \frac{3-\sqrt{5}}{2}$. In item (2), because $\lambda > \frac{1}{2}$, we must have $A < 0$; thus $4\lambda^2 A - (2\lambda-1)^2 < 0$, and $\Pi_c^* < \Pi_{nc}^*$. In items (3)–(6), $\Pi_c^* > \Pi_{nc}^*$ is equivalent to the condition that $\alpha < \bar{\alpha}(\lambda) \in (0,1)$. ■

## Proof of Proposition 5

First, we solve the case for uninformed consumers.

Denote an uninformed consumer's valuation as $CS_{nc}^u$ and $CS_c^u$ for when the firm does not collect data or collects data, respectively. By Proposition 3, without data collection, her expected surplus $CS_{nc}^u$ equals 0 if $\alpha \leq \frac{\lambda+\lambda^2}{1+\lambda^2}$ and $\frac{\lambda(1-\alpha)}{1-\lambda}\cdot[(1+\lambda^2)\alpha - (\lambda+\lambda^2)]$ if otherwise. But when the firm collects data, she can infer her true type; thus her valuation will be updated according to Proposition 1. Because she assigns a conditional probability $\lambda$ to be the uninformed $H$-types, her expected surplus with data collection $CS_c^u$ equals $\frac{\lambda}{2}(1-\lambda)$ when $\alpha \leq \lambda$ and equals $\frac{\lambda(1-\alpha)}{2(1-\lambda)}(1+\alpha-2\lambda)$ when $\alpha > \lambda$. Clearly, $CS_{nc}^u = 0 < CS_c^u$ whenever $\alpha \leq \frac{\lambda+\lambda^2}{1+\lambda^2}$. Because $\alpha > \frac{\lambda+\lambda^2}{1+\lambda^2}$ implies $\alpha > \lambda$, it suffices to compare $\frac{\lambda(1-\alpha)}{1-\lambda}[(1+\lambda^2)\alpha - (\lambda+\lambda^2)]$ and $\frac{\lambda(1-\alpha)}{2(1-\lambda)}(1+\alpha-2\lambda)$. Because $\alpha < 1$, we have $CS_{nc}^u < CS_c^u$ if $\alpha > \frac{\lambda+\lambda^2}{1+\lambda^2}$. Combining both results, we conclude that $CS_{nc}^u < CS_c^u$ regardless of $\alpha$.

Second, we derive the conditions in which data collection benefits average consumers by checking the following conditions:

(1) If $\alpha > \frac{\lambda+\lambda^2}{1+\lambda^2}$, then
$\frac{\alpha_H + 3\alpha_L - 4\lambda\alpha_H}{2(1-\lambda)} \geq 2\alpha_L - \frac{\lambda(1+\lambda+3\lambda^2-\lambda^3)(\alpha_H-\alpha_L)}{1-\lambda}$.

(2) If $\frac{\lambda+\lambda^2}{1+\lambda^2} \geq \alpha > \lambda$, then
$\frac{\alpha_H + 3\alpha_L - 4\lambda\alpha_H}{2(1-\lambda)} \geq (1-\lambda)(\alpha_L + \lambda^2(\alpha_H - \alpha_L))$.

(3) If $\alpha \leq \lambda$, then
$\frac{\alpha_H^2}{2} \geq (\alpha_H - \alpha_L)(\alpha_L + \lambda^2(\alpha_H - \alpha_L))$.

Note that condition (1) is equivalent to $(1-\lambda)^2 + \lambda^2(1-\lambda)(1+2\lambda) + 5\lambda^3 \geq 0$. Thus for any $\alpha > \frac{\lambda+\lambda^2}{1+\lambda^2}$, $CS_c^* \geq CS_{nc}^*$. Condition (2) is equivalent to $2(1-\lambda^2) + (\alpha-1)[3 - 2\cdot(1-\lambda^2)(1-\lambda)^2] \geq 0$. Because $\alpha > \lambda$, and $3 - 2(1-\lambda^2)\cdot(1-\lambda)^2 \geq 3 - 2 > 0$, we have $2(1-\lambda^2) + (\alpha-1)[3 - 2(1-\lambda^2)\cdot(1-\lambda)^2] > 2(1-\lambda^2) + (\lambda-1)[3 - 2(1-\lambda^2)(1-\lambda)^2] = (1-\lambda)\cdot[(1-\lambda^2)(1-\lambda)^2 + \lambda^3(2-\lambda)] > 0$. Therefore, for any $\alpha > \lambda$, $CS_c^* \geq CS_{nc}^*$. Finally, condition (3) holds if and only if either of the two following conditions holds:

(i) $\lambda < \frac{\sqrt{2}}{2}$ or

(ii) $\lambda \geq \frac{\sqrt{2}}{2}$ and $\alpha \geq \frac{\sqrt{2\lambda^2-1} - (2\lambda^2-1)}{2(1-\lambda^2)}$.

With $\alpha \leq \lambda$, it remains to show that $\frac{\sqrt{2\lambda^2-1} - (2\lambda^2-1)}{2(1-\lambda^2)} \leq \lambda$, which is equivalent to showing that if $\alpha = \lambda$, then $\frac{1}{2} \geq (1-\alpha)\cdot[\alpha + \lambda^2(1-\alpha)]$. Substitute $\alpha = \lambda$ in the inequality to obtain that $(1-\lambda^2)(1-\lambda)^2 + \lambda^3(2-\lambda) \geq 0$. Therefore, $\frac{\sqrt{2\lambda^2-1} - (2\lambda^2-1)}{2(1-\lambda^2)} \leq \lambda$.

In summary, the average consumer surplus is higher with data collection if and only if

(i) $\lambda < \frac{\sqrt{2}}{2}$ or

(ii) $\lambda \geq \frac{\sqrt{2}}{2}$ and $\alpha \geq \frac{\sqrt{2\lambda^2-1} - (2\lambda^2-1)}{2(1-\lambda^2)}$. ■

## Proof of Proposition 6

Under the conditions of Proposition 4, the firm benefits strictly from data collection. The $L$-types receive negative surplus without data collection but positive surplus with. We now turn to $H$-types' surplus, which is given in Table A.1. From Table A.1, $H$-types obtain higher expected surplus with data collection if and only if any of the following conditions holds:

(1) $\alpha + \lambda^2(1-\alpha) < \frac{1}{2(1-\alpha)}$ if $\alpha \in (0, \lambda]$,

(2) $\alpha + \lambda^2(1-\alpha) < \frac{1+3\alpha-4\lambda}{2(1-\lambda)^2}$ if $\alpha \in (\lambda, \frac{\lambda+\lambda^2}{1+\lambda^2}]$, or

(3) $\alpha(1 - \lambda + 3\lambda^3 - \lambda^4) - \lambda^3(3-\lambda) < \frac{1+3\alpha-4\lambda}{2}$ if $\alpha \in \left(\frac{\lambda+\lambda^2}{1+\lambda^2}, 1\right)$.

First, $\alpha + \lambda^2(1-\alpha) < \frac{1}{2(1-\alpha)}$ is equivalent to $(1-\alpha)^2\cdot(1-2\lambda^2) + \alpha^2 > 0$. Therefore, condition (1) holds if and only if either (a) $1 - 2\lambda^2 \geq 0$ or (b) $1 - 2\lambda^2 < 0$ and $\alpha > \frac{1-2\lambda^2+\sqrt{2\lambda^2-1}}{2-2\lambda^2}$. Second, condition (2) is equivalent to $\alpha > \frac{2(1-\lambda)^2\lambda^2+4\lambda-1}{3-2(1-\lambda)^2(1-\lambda^2)}$, which is implied by $\alpha > \lambda$. To see this, we will show $\lambda > \frac{2(1-\lambda)^2\lambda^2+4\lambda-1}{3-2(1-\lambda)^2(1-\lambda^2)}$: because $3 - 2(1-\lambda)^2(1-\lambda^2) > 0$, reorganize the inequality to obtain that $[(2-3\lambda)^2 + 3\lambda^2]\cdot[1 + \lambda(1-\lambda)] + \lambda^2(1-\lambda)^2 > 0$, which is apparent because $\lambda \in (0,1)$, and thus condition (2) holds. Third, condition (3) is equivalent to $2\lambda^4 - 6\lambda^3 + 2\lambda + 1 - \frac{2(1-\lambda)}{1-\alpha} < 0$. If $\alpha > \lambda$, then $-\frac{2(1-\lambda)}{1-\alpha} < -2$; thus $2\lambda^4 - 6\lambda^3 + 2\lambda + 1 - \frac{2(1-\lambda)}{1-\alpha} < 2\lambda^4 - 6\lambda^3 + 2\lambda - 1 = -2\lambda(1-\lambda)^3 - (1-2\lambda)^2 - 2\lambda^2 < 0$. Therefore, condition (3) holds. Consequently, $H$-types obtain higher

expected surplus with data collection if and only if either of the following conditions holds: $\lambda < \frac{\sqrt{2}}{2} \approx 0.707$ or $\alpha > \underline{\alpha} \equiv \frac{1-2\lambda^2+\sqrt{2\lambda^2-1}}{2-2\lambda^2}$. Because $\lambda < \frac{(3-\sqrt{5})}{2} = 0.382 < 0.707$ always holds (under the conditions of Proposition 4), the condition $\alpha > \underline{\alpha}$ is necessary and sufficient for $H$-types to be better off with data collection. ■

**Table A.1.** *H*-Types' Ex Post Surplus

| | | $\alpha \in (0, \lambda]$ | $\alpha \in (\lambda, \frac{\lambda+\lambda^2}{1+\lambda^2}]$ | $\alpha \in (\frac{\lambda+\lambda^2}{1+\lambda^2}, 1)$ |
|---|---|---|---|---|
| No data collection | $m \in \mathcal{M}^0 \cup \mathcal{M}^1$ | $(1-e)[\alpha + \lambda^2(1-\alpha)]$ | | $(1-\alpha)\left[\alpha - \frac{\lambda^3(3-\lambda)}{1-\lambda}(1-\alpha)\right]$ |
| Data collection | $m \in \mathcal{M}^0$ | 0 | $\frac{(\alpha-\lambda)(1-\alpha)}{(1-\lambda)}$ | |
| | $m \in \mathcal{M}^1$ | $\frac{(1-\lambda)}{2}$ | $\frac{(1+\alpha-2\lambda)(1-\alpha)}{2(1-\lambda)}$ | |

## Endnotes

[1] In fact, qualitative research techniques, such as observational methods, are used by marketers precisely to uncover what consumers themselves do not know about their preferences. For instance, instead of asking consumers, marketers for Pernod Ricard USA observed consumers in their own homes hosting parties to understand how attendees interacted with various brand-named beverages. (See https://www.theatlantic.com/magazine/archive/2013/03/anthropology-inc/309218/, accessed February 2018.) In addition, Larréché (2008) discussed situations in which firms may design new products even before consumers learn to value the innovating features.

[2] See "Big Data & Differential Pricing," February 2015, from the U.S. President's Council of Economic Advisers, obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf (accessed June 4, 2019).

[3] By contrast, whereas our research considers consumers' discomfort about the marketer's use of data to take advantage of them in the marketplace, it does not account for their privacy concerns of identity theft and personalized price discrimination.

[4] The model generalization is provided in the online appendix.

[5] See Hummel et al. (2013) about the incentives of consumers to misreport their perceptions. We take as given that consumers report their signals truthfully.

[6] To see this why this assumption suffices, suppose that $m > m'$, without loss of generality. Then by the strict monotonicity property, $\theta_i(m) > \theta_i(m')$ for any $i = H, L$. Consider $i \neq j$, and $\mathbf{\Theta}(m) = \{\theta_i(m), \theta_j(m)\}$. Because $\theta_i(m) = \theta_j(m')$ implies $\theta_j(m) > \theta_j(m') = \theta_i(m) > \theta_i(m')$, we must have $\mathbf{\Theta}(m) \neq \mathbf{\Theta}(m')$ for any $m \neq m'$.

[7] Please see the general proof in the online appendix for more details.

[8] One may be interested to assess the impact of superior knowledge when there is also uncertainty on the distribution of the consumer types ($\lambda$) or the intrinsic values ($\alpha$). The uncertainty on the proportion of consumer types ($\lambda$) has been analyzed by Taylor (2004). As for the uncertainty on $\alpha_i$, we explore the possibility that each type's value is subject to random noise. An analysis of this case, which is relegated to the online appendix, illustrates that the firm can acquire even greater informational advantage over the consumers. In particular, the firm acquires a more precise estimate of $\alpha_i$ than not only uninformed consumers but also informed consumers who learn their type.

[9] We assume that consumers are not allowed to share their $\theta$ with other consumers. Therefore, observing two data points of different consumer types is enough for the firm to acquire superior knowledge. But even if consumers can share their data with each other, superior knowledge is still possible whenever there are many consumers of each type and the firm acquires more data than at least some consumers could acquire through sharing.

[10] Note that $\mathrm{BR}(\Psi', \mu_i)$ depends on the belief $\mu_i$, rather than directly on $m$, which the consumer does not observe.

[11] Formal details of the equilibrium refinement process are relegated to the proof of Proposition 1 in the appendix.

[12] Formally, $\mu_H$ survives D1 if and only if for any out-of-equilibrium pricing strategy $\Psi' \neq \Psi^*$, $i \in \{H, L\}$, and $j \neq k \in \{0, 1\}$; then $\mu_H(\Psi') = 1$ if $k = 1$ and $\mu_H(\Psi') = 0$ if $k = 0$, whenever the following condition holds: $\cup_\mu\{r_H | \Pi^*(\mathcal{M}^j) \leq \hat{\Pi}_H(\mathcal{M}^j, \Psi', r_H)\} \subsetneq \cup_\mu\{r_H | \Pi^*(\mathcal{M}^k) < \hat{\Pi}_H(\mathcal{M}^k, \Psi', r_H)\}$, where $\cup_\mu\{.\}$ is a set of arbitrary beliefs $\mu \in [0,1]$ such that the condition inside the braces holds, and $r_i$ is a best response to $\Psi'$ given the arbitrary belief $\mu$; that is, $r_H = \mathrm{BR}(\Psi', \mu \in [0,1])$.

[13] Similarly, even though the firm with collected data may have incentives to facilitate consumer communication to alleviate the signaling costs, strategic consumers may be reluctant to share their private realizations of $\theta_i$ with other consumers.

[14] For reference, see Xu and Dukes (2019).

## References

Acquisti A, Varian HR (2005) Conditioning prices on purchase history. *Marketing Sci.* 24(3):367–381.

Anderson ET, Dana JD (2009) When is price discrimination profitable? *Management Sci.* 55(6):980–989.

Anderson LR, Holt CA (1997) Information cascades in the laboratory. *Amer. Econom. Rev.* 87(5):847–862.

Banks JS, Sobel J (1987) Equilibrium selection in signaling games. *Econometrica* 55(3):647–661.

Bettman JR, Luce MF, Payne JW (1998) Constructive consumer choice processes. *J. Consumer Res.* 25:187–217.

Bettman JR, Luce MF, Payne JW (2008) Preference construction and preference stability: Putting the pillow to rest. *J. Consumer Res.* 18(3):170–174.

Calzolari G, Pavan A (2005) On the optimality of privacy in sequential contracting. *J. Econom. Theory* 30(1):168–204.

Cho I-K, Kreps DM (1987) Signaling games and stable equilibria. *Quart. J. Econom.* 102(2):179–221.

Cho I-K, Sobel J (1990) Strategic stability and uniqueness in signaling games. *J. Econom. Theory* 50(2):381–413.

Fudenberg D, Villas-Boas JM (2006) Behavior-based price discrimination and customer recognition. Hendershott T, ed. *Economics and Information Systems*, Handbooks in Information Systems, vol. 1 (Emerald Group Publishing, Bingley, UK), 377–436.

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

Hummel P, Morgan J, Stocken P (2013) A model of flops. *RAND J. Econom.* 44(4):585–609.

Kamenica E (2008) Contextual inference in markets: On the informational content of product lines. *Amer. Econom. Rev.* 98(5): 2127–2149.

Kamenica E, Mullainathan S, Thaler R (2011) Helping consumers know themselves. *Amer. Econom. Rev.* 101(3):417–422.

Larréché JC (2008) *The Momentum Effect: How to Ignite Exceptional Growth* (Wharton School Publishing, Upper Saddle River, NJ).

Milgrom PR, Weber RJ (1985) Distributional strategies for games with incomplete information. *Math. Oper. Res.* 10(4):619–632.

Moorthy KS (1984) Market segmentation, self-selection, and product line design. *Marketing Sci.* 3(4):288–307.

Mussa M, Rosen S (1978) Monopoly and product quality. *J. Econom. Theory* 18(2):301–317.

Roettgers J (2018) Netflix's secrets to success: Six cell towers, dubbing and more. *Variety.com* (March 8), https://variety.com/2018/digital/news/netflix-success-secrets-1202721847/.

Samuelson P (2000) Privacy as intellectual property. *Stanford Law Rev.* 52(5):1125–1173.

Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. *J. Risk Uncertainty* 1(1):7–59.

Simonson I (2008) Will I like a "medium" pillow? Another look at constructed and inherent preferences. *J. Consumer Behav.* 18(3):155–169.

Solove DJ (2007) "I've got nothing to hide" and other misunderstandings of privacy. *San Diego Law Rev.* 44:745–772.

Stephens-Davidowitz S (2017) *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are* (HarperCollins, New York).

Stokey NL (1979) Intertemporal price discrimination. *Quart. J. Econom.* 93(3):355–371.

Taylor CR (2004) Consumer privacy and the market for customer information. *RAND J. Econom.* 35(4):631–650.

Varian HR (1985) Price discrimination and social welfare. *Amer. Econom. Rev.* 75(4):870–875.

Varian HR (2002) Economic aspects of personal privacy. Lehr WH, Pupillo LM, eds. *Cyber Policy and Economics in an Internet Age* (Kluwer Academic Publishers, Norwell, MA), 127–137.

Villas-Boas JM (2004) Price cycles in markets with customer recognition. *RAND J. Econom.* 35(3):468–501.

Wernerfelt B (1995) A rational reconstruction of the compromise effect: Using market data to infer utilities. *J. Consumer Res.* 21(4): 627–633.

Xu Z, Dukes AJ (2019) Personalized pricing with superior preference information and the role of list price. Working paper, University of Southern California, Los Angeles.