How (not) to pay for advice: A framework for consumer financial protection*

Roman Inderst† Marco Ottaviani‡

June 2011

Abstract

This paper investigates the determinants of the compensation structure for brokers who advise customers regarding the suitability of financial products. Our model explains why brokers are commonly compensated indirectly through contingent commissions paid by product providers, even though this compensation structure may lead to biased advice. We investigate when policy intervention that mandates disclosure or prohibits commissions can protect naive consumers from exploitation or enhance efficiency by leading to more informative advice.

Journal of Economic Literature Classification Codes: D18, D83, G24, G28.

Keywords: Brokers, financial advisers, commissions, consumer financial protection, disclosure.

*We thank seminar participants at UCLA, University of Chicago Booth School of Business, Michigan State University, University of Mannheim, University of Rochester Simon Graduate School of Business, and University of Vienna, audiences at the FTC-Northwestern Microeconomics Conference 2009, the NBER Law and Economics Summer Institute 2010, the Utah Winter Business Economics Conference 2010, and the AEA 2011 Meetings, and discussants Heski Bar-Isaac, Boğaçhan Çelen, Oliver Hart, and Kristóf Madarász.

†Johann Wolfgang Goethe University Frankfurt (IMFS) and Imperial College London. E-mail: inderst@finance.uni-frankfurt.de.

‡Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208-2013, USA. E-mail: m-ottaviani@northwestern.edu.

"Impartial advice represents one of the most important financial services consumers can receive. . . . Mortgage brokers often advertise their trustworthiness as advisers on difficult mortgage decisions. When these intermediaries accept side payments from product providers, they can compromise their ability to be impartial. Consumers, however, may retain faith that the intermediary is working for them and placing their interests above his or her own, even if the conflict of interest is disclosed. Accordingly, in some cases consumers may reasonably but mistakenly rely on advice from conflicted intermediaries." *Financial Regulatory Reform. A New Foundation: Rebuilding Financial Supervision and Regulation, US Department of Treasury, June 2009 (page 68)*

1 Introduction

Across countries, customers rely on recommendations from brokers and other financial advisers when making important decisions about purchasing financial services such as mortgages, consumer credit, life insurance, and investment products.[1] In many instances, however, the recommendations may be biased, because often the advising intermediaries are not paid directly by customers but, instead, receive commissions and other distribution fees from the providers of financial products.[2] These payments may tilt their recommendations toward particular financial products.[3] Likewise, when the payments from product providers are proportional to the size of transactions (or when the adviser is compensated only when a transaction is made), customers may be induced to take larger positions (or

[1] A large-scale survey conducted in 2003 by the European Commission (Eurobarometer 60.2, November-December 2003) documents that in many European countries such as Finland, Germany, and Austria more than 90% of respondents expect to receive advice from financial institutions. Also for the US, the role of professional financial advice for the purchase of investment products (outside employer-sponsored plans) has been much documented. For instance, in a survey conducted by the Investment Company Institute (ICI 2007), over 80 percent stated that they obtained financial advice from professional advisors or other sources (cf. also *Equity Ownership in America* 2005, http://www.ici.org/pdf/rpt_05_equity_owners.pdf).

[2] According to a pool of the EU members of the CFA Institute (2009), 64% of respondents "believe that the fee structure of investment products drives their sale to customers rather than their suitability to customers."

[3] "Many borrowers whose credit scores might have qualified them for more conventional loans say they were pushed into risky subprime loans. . . . The subprime sales pitch sometimes was fueled with faxes and emails from lenders to brokers touting easier qualification for borrowers and attractive payouts for mortgage brokers who brought in business. One of the biggest weapons: a compensation structure that rewarded brokers for persuading borrowers to take a loan with an interest rate higher than the borrower might have qualified for." Subprime Debacle Traps Even Very Credit-Worthy As Housing Boomed, Industry Pushed Loans To a Broader Market, *Wall Street Journal*, December 3, 2007.

to make more frequent transactions).[4]

There is growing concern among consumer groups and government regulators that indirect compensation based on commissions may lead to unsuitable advice.[5] Would customers of retail financial services be better served if, instead, intermediaries were paid directly, through an hourly fee?[6] Brokers or financial advisers would then earn the same compensation regardless of the ultimate decision of the customer and would thus no longer be biased toward recommending a particular product or service. But if the prevalent compensation structure for advice seriously compromises its value, why would intermediaries and product providers not find a more efficient arrangement?[7]

This paper offers a rationale for the prevailing compensation structure and then investigates the need for policy intervention from a normative perspective. To this end, we propose a model in which customers vary in their understanding of the advisers' conflict of interest. While wary customers understand that product providers have incentives to pay commissions to advisers to steer customers toward their offerings, naive customers believe that advisers are unbiased. The model jointly endogenizes the payments that product providers make to intermediaries such as financial advisers as well as the way customers pay for financial products and advice. In equilibrium, lower up-front fees for advice but higher product prices (in the form of higher loads for investment products or higher interest rates on loans) are associated with higher commissions or other inducements that are paid to advisers or brokers.

To set the stage, consider the benchmark case in which customers are wary about the

[4]Hackethal, Inderst, and Meyer (2010) document how branches of a large German bank make considerably higher revenues from increased security transactions when retail customers report to strongly rely on the bank's advice.

[5]The European Commission has singled out the provision of precontractual information through advice as one of the three main problem areas for the retail financial sector. In particular, see pages 12–14 of the staff working document of the Commission of the European Communities (2009).

[6]In a recent consultation document, the UK financial regulator Financial Services Authority (2009), henceforth FSA, has proposed steps to encourage a *complete* switch toward a regime in which customers pay independent financial advisers directly. The new rules would "require adviser firms to be paid by adviser charges: the rules do not allow adviser firms to receive commissions offered by product providers." As part of a package of sweeping reforms enacted in the wake of the financial crisis, the US Consumer Financial Protection Act of 2010 has instituted a Bureau of Consumer Financial Protection which has authority to write such rules to protect consumers; see the Dodd–Frank Wall Street Reform and Consumer Protection Act, Title X.

[7]Bergstresser, Chalmers, and Tufano (2007) and Chen, Hong, and Kubik (2007), for instance, suggest that mutual funds sold through broker/agent networks tend to underperform and that funds with higher fees improve distribution through higher commissions.

advisers' incentives. We show that even when financial inducements to advisers can be paid secretly, there need not be a commitment problem vis-à-vis wary customers, provided that contracts are sufficiently flexible. Precisely, we show how this result holds when, first, contracts are sufficiently flexible to overcome an agency problem between product providers and advisers and, second, consumer surplus can be extracted through a fixed fee for advice. Formally, the first condition is verified when advisers are not wealth constrained and can thus transfer profits to product providers through a lump-sum payment. More generally, as we discuss below, the notion of joint-profit maximization may be more applicable when product providers and advisers engage in a long-term relationship, rather than operating at arm's length. With wary customers, joint profits are maximized when a product provider credibly commits *not* to pay secret inducements that bias advice. Intuitively, this outcome is achieved when a low price is charged for the product. Advisers then charge wary consumers a high fixed fee, which in turn is transferred to product providers.

In equilibrium wary customers would rationally anticipate how a higher price that they pay to the product provider is passed through into higher commissions to intermediaries, and how these commissions ultimately affect recommendations and choices. However, when some customers, as in the opening quote from the US Treasury, naively fail to adequately take into account the potentially self-interested nature of advice, the fee structure that prevails in equilibrium is no longer efficient. As we show, product providers are able to better exploit the misperceptions of naive customers by inducing a compensation structure involving a lower up-front charge for advice and a higher final price. In fact, when all customers are naive in this way, our model predicts that customers are not asked to pay any up-front charges for advice. Then, intermediaries are only compensated indirectly through the commission payments they receive from product providers.

In equilibrium, naive customers underestimate the likelihood with which they end up purchasing an "advanced" premium product (or a product at all) that generates higher profits for the respective financial institutions and for the intermediary than a more "basic" offering (or no purchase). Even though customers appear not to pay for advice, in reality they are thus seriously shortchanged through biased advice and higher product prices, in the form of higher management fees on investment products or higher interest rates on mortgages.

With naive customers, there is then a clear benefit of policy intervention that requires

firms to make customers pay directly for advice. A cap (or, ultimately, a ban) on commissions or other inducements increases consumer surplus by restricting the extent to which the customers' naive beliefs can be exploited. With a mixed population of wary and naive customers, policy intervention also affects the incentives of product providers to target different segments of the population. In fact, in the absence of policy intervention, when the market is populated mostly by naive customers, firms may generate higher profits by targeting exclusively naive customers rather than serving the whole market with a non-exploitative offer.

Policy intervention can, however, backfire when the practice of paying "indirectly" for advice arises in the presence of wary customers, who see through the incentives of financial institutions and intermediaries. With wary customers, we highlight an efficiency rationale for compensating intermediaries also through commissions paid by product providers. Even though indirect pay for advice leads to biased advice, the overall quality of advice that results may be higher because the adviser's incentives to acquire information are improved. It may thus be efficient *not* to perfectly align the interests of advisers with those of wary customers at the recommendation stage. Specifically, even when customers are wary of the conflict of interest and presence of commissions, we show that high commissions result for products that are likely ex ante to suit the preferences and needs of only a small fraction of customers. Intuitively, in the absence of commissions, the adviser would have little incentive to learn whether such products are indeed suitable for a customer. For relatively more complex and specialized products, for which it is optimal that the adviser be better informed, capping or prohibiting commissions may thus have more severe unintended consequences.

These negative side effects of hard-handed policy intervention can be avoided with a policy of mandatory disclosure of financial inducements paid to advisers, provided that disclosure turns otherwise naive customers into wary—which is why firms themselves may be reluctant to provide such information.[8] In fact, we show that customer naiveté dampens competition and leads to higher joint profits in the long run, so that even with competition firms may have little incentives to educate customers.

[8]In the US, the Federal Trade Commission (2008) has proposed rules that would require that brokers enter in an initial agreement with customers that "must state that the consumer will pay the entire compensation even if all or part is paid directly by the lender, and that a lender's payment to a broker can influence the broker to offer the consumer loan terms or products that are not in the consumer's interest or are not the most favorable the consumer could obtain."

Our exploitation result is reminiscent of DellaVigna and Malmendier (2004). While in their model customers are naive about their future demand, in our model customers are naive about the incentives behind the advice received.[9] Given that incentives are endogenously determined in our model, firms exploit the customers' naiveté by increasing the conflict of interest through commissions. To what extent can customers be expected to be sufficiently wary of the conflict of interest when their advisers are paid through commissions or other inducements? The form of naiveté about incentives that we posit is similar to the one documented empirically by Malmendier and Shanthikumar (2007) in the context of recommendations made by security analysts to investors.[10] Using data from the Survey of Consumer Finances, Bergstresser and Beshears (2010) show that borrowers who were less able to comprehend financial questions and who were less suspicious in interviews were more likely to purchase adjustable-rate mortgages in the period 2004–2007; these mortgages then exhibited higher rates of foreclosure than fixed rate mortgages during the mortgage crisis. Chater, Inderst, and Huck (2010) show in a survey among six thousand recent purchasers of retail financial services in Europe that respondents are largely ignorant of conflicts of interest and, indeed, rarely pay directly for advice.[11]

To the fledgling literature on consumer financial protection, we contribute a positive and normative analysis of the compensation structure for advice. Other recent contributions in the area focus on different aspects relevant to the provision of non-verifiable information to customers.[12] Bolton, Freixas, and Shapiro (2007) analyze how incentives for information provision depend on competition among banks. Inderst and Ottaviani (2009) focus on the multi-task agency problem a seller faces when hiring an agent to find as well as to advise customers. Inderst and Ottaviani (2010a) analyze competition through commissions as well as through prices among multiple product providers in a common agency

[9] Also Carlin (2009) considers customers with varying degrees of sophistication; in his model, however, sophisticated customers are able to observe individual prices, while non-sophisticated customers purchase randomly.

[10] Experiments with games of trust and cheap talk also suggest that many subjects are willing to follow advice more than they should, even when payoffs and incentives are revealed (e.g., Cain, Loewenstein, and Moore, 2005).

[11] In particular, more than half of the respondents thought that financial advisors or the staff of a tied provider gave completely independent advice or information. Only a minority believed or even knew that the intermediary through which they purchased a product received a commission or a bonus for selling the investment. Of those purchasing through a financial adviser or a broker, only around 5% reported to have paid a direct fee for advice.

[12] Earlier papers, such as Admati and Pfleiderer (1986), analyze how a seller should optimally charge for information when its quality can be verified by customers.

framework.

In an early contribution cast in the context of insurance markets, Gravelle (1994) also analyzes the compensation structure of brokers. In Gravelle's (1994) model, however, brokers truthfully reveal to customers the valuation for the product, so that the choice between up-front payment and commission trades off two monopoly-pricing problems; the up-front payment reduces the number of customers who become informed, whereas the commission charge reduces the number of informed customers who actually purchase the insurance product. Gravelle (1993) captures the activity of insurance brokers with respect to unsophisticated customers through an upward shift in demand. In a similar vein, Stoughton, Wu, and Zechner (2011) analyze how intermediaries can be incentivized to market more aggressively investment products to unsophisticated investors. In their analysis of delegated investment management, kickbacks paid by portfolio managers to intermediaries enable investment fund managers to price discriminate across investors with more or less wealth.

The paper proceeds as follows. Section 2 introduces the baseline model. Section 3 analyzes the provision of advice. Sections 4, 5, and 6 solve for the equilibrium compensation structure and advice in the presence of wary customers, naive customers, and a heterogeneous population with both types of customers. Sections 7, 8, and 9 extend the model to analyze the effect of agency frictions, competition, and endogenous information acquisition. Section 10 summarizes the policy implications. Section 11 concludes. Appendix A collects the proofs of all the propositions reported in the paper. Appendix B analyzes an analytical example.

2 Baseline model

We are interested in analyzing some generic features of the market for many retail financial services, such as investment products, pension plans, mortgages, and life insurance policies. Abstracting from specific features of markets for particular products and services, we frame our analysis more generally in terms of a customer's choice between two options. This choice is based on an adviser's recommendation regarding the suitability of the characteristics of either option to the customer's specific needs and circumstances, such as the customer's wealth, earnings prospects, age, risk attitude, and tax status. For instance, the attractiveness of a fixed rate mortgage (FRM) relative to an adjustable rate

mortgage (ARM) depends on the indexation of the borrower's income stream. A household's optimal choice of pension scheme, in terms of risk and liquidity, depends on factors such as age to retirement and risk tolerance given the composition of the household's asset portfolio.[13] Similarly, the tax implications of different investment vehicles, such as stocks and municipal bonds, depend on an investor's tax bracket.

Products, customer preferences, and advice. As represented schematically in Figure 1, we denote the customer's options by $\theta = A, B$, where A corresponds to the choice of product A, while B may stand for another product or, alternatively, for the option of not purchasing at all. Our analysis applies to both cases. In case the two options correspond to different products, we may think of B as representing the "basic" (or default) option, while A represents the "advanced" (or premium) option. For instance, option B may represent the option of not investing or that of investing in Treasury bills, while option A may represent a mutual fund or a structured product. Alternatively, B could be a plain vanilla mortgage (such as an FRM) and A a more innovative arrangement.

The price of product A, $p_A = p$, is chosen by the respective product provider. It may represent management fees or required interest payments. To focus our analysis, in our baseline specification we assume that the payoff of the alternative option, B, is exogenously given. In case B represents an alternative product, rather than the option of not purchasing at all, then we may suppose that its price p_B is determined competitively, and thus it is equal to cost. To streamline the notation we set equal to zero all costs, i.e., both the cost of providing each product and the cost of administering a purchase. Thus, in this baseline setting there are three strategic players: the monopolistic provider of the advanced product A, the adviser, and the customer.

The value realized by the customer depends on the match between the customer's preferences and needs with the characteristics of the options available. We capture the importance of the match by supposing that there are two customer types, $\widehat{\theta} = A, B$, with corresponding utilities $v_{\theta,\widehat{\theta}}$ in case product θ is matched with customer type $\widehat{\theta}$. The key assumption is that a fitting match creates higher utility, $v_{A,A} > v_{B,A}$ and $v_{B,B} > v_{A,B}$. We impose symmetry by supposing that $v_{A,A} = v_{B,B} = v_h$ and $v_{A,B} = v_{B,A} = v_l$, with $v_h > v_l$,

[13] In a recent review of the advice provision for personal pension plans, the UK Financial Service Authority (2010) reported many instances of advised pension switches that were unsuitable given customers' attitude to risk, often in addition to involving an inappropriate loss of benefits from the ceding scheme.

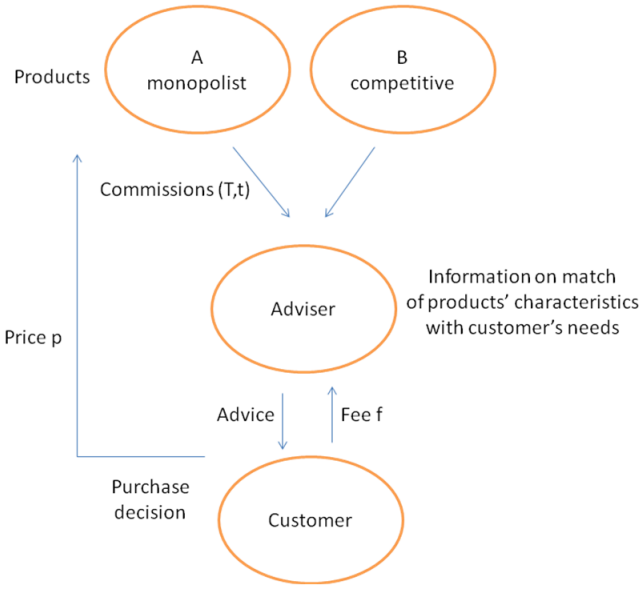

Figure 1: **Scheme for the baseline model.**

and we define $\Delta_v := v_h - v_l$.

The initial (or prior) public probability that choice A is more suitable is equal to q_0. The customer's expected gross payoff is then $v_l + q_0\Delta_v$ when choosing A, and it is $v_l + (1 - q_0)\Delta_v$ when choosing B. We assume that the basic option is more suitable for the average customer, $q_0 < 1/2$, so that the advanced option constitutes a niche market.

The customer can obtain advice from an adviser who acts as information intermediary. Presently, we consider the quality of the adviser's information to be exogenously given. This information is captured by the cumulative distribution function of the adviser's posterior belief, $G(q)$, with full support $q \in [0, 1]$.[14] By Bayesian updating the expected posterior belief is equal to the prior, so that $\int_0^1 [1 - G(q)]\, dq = q_0$. In Section 9 we consider costly information acquisition by the adviser, which is then modeled by a transformation of $G(q)$. Also, this baseline model features a single customer demanding a single unit of the product sold by a monopoly network comprising a provider and an adviser. In Section 8 we extend the model to allow for a downward-sloping demand for products and to analyze competition between networks.

[14] Even though it is convenient to take the distribution of posterior beliefs as the primitive, clearly this distribution can be generated by Bayesian updating from an underlying private signal s that the adviser observes with conditional distributions $H_A(s)$ and $H_B(s)$.

Contracting between product providers and adviser. Consider first the contract between product provider A and the adviser. In our baseline model, the contract prescribes two elements, a fixed payment T and a conditional payment t that is paid only when subsequently product A is sold. Presently, we also do not place any sign restrictions on T, which thus can be set to be negative so as to eliminate the internal agency problem in the distribution chain. This baseline scenario without agency frictions allows us to focus on the contracting problem with respect to customers. We analyze in Section 7 the case in which the agency problem may not be resolved perfectly because of the constraint that $T \geq 0$.

Note also that the adviser does not receive additional payment when option B is chosen, so that either no purchase is made or the basic (and competitively provided) product is purchased. It is, however, straightforward to extend the analysis to allow for payments that would need to be made to the adviser in order to cover any administrative or handling costs. After all, what will matter for our analysis is the *difference* between the payments that the adviser receives when the customer makes the respective choices.

The contingent payment t may take different forms in practice. For some investment products, the broker or independent financial adviser may receive all or a fraction of the "load" that the customer initially pays to the product provider. More generally, the intermediary may receive a commission. With credit products, brokers' compensation is often tied to the interest rate through the so-called "yield spread;" see Jackson and Burlingame (2007). Sellers of life insurance plans may be paid both up-front or via a trail-commission over the duration of the contract; see Cummins and Doherty (2006).

When making a recommendation, the adviser is also concerned about the suitability of the option chosen by the customer. We capture this concern by stipulating that the adviser's future payoff is reduced by $\rho > 0$ when the customer ultimately realizes low utility v_l instead of high utility v_h. Even though the respective levels of the adviser's payoff is inessential for our analysis, for concreteness we specify that the adviser's payoff, gross of payments received from product providers, is equal to u_l when v_l is realized and equal to u_h when v_h is realized, so that $\rho = u_h - u_l$. This simple way of modeling the suitability concern follows Bolton, Freixas, and Shapiro (2007) and Inderst and Ottaviani (2009).[15]

[15] Bolton, Freixas, and Shapiro (2007) and Inderst and Ottaviani (2010a) also show how to endogenize ρ in a dynamic model where the penalty is due to the loss of future business following an unsuitable sale. In Inderst and Ottaviani (2010b) such a penalty arises from the contractually stipulated cancellation terms

By specifying that also $u_l \geq 0$, we can restrict the adviser's recommendations to either A and B even when option B represents an alternative, more basic product, so that, in principle, the third option of recommending not to purchase is also present.

The adviser's concern for suitability may have different origins. The adviser may simply have professional concerns about a customer's well-being. There may also be reputational costs, e.g., through the loss of future business with this or other customers. Further, ρ may capture the prospect of prosecution by courts or regulators following customer complaints regarding suitability or a review of past sales by supervising authorities. To be specific, we suppose that ρ represents a fine paid to regulators.[16]

Contracting with customers. When purchasing product A, customers must pay the respective price $p_A = p$. Recall that the payment that must be made when choosing option B is set to zero. As discussed in the Introduction, we allow the adviser to stipulate a flat fee f for advice; this is a key innovation of our analysis. We restrict this fee to be nonnegative, $f \geq 0$ according to a standard "no free lunch" condition that prevents the adviser from bribing the customer into business with a positive up-front payment. A standard assumption to rule out such up-front transfers is the presence of a sufficiently large pool of frivolous customers, who would then turn up to cash in the fixed payment while having no intention to make a purchase. Only when the adviser's offer is accepted by a customer, who arrives next, does the game proceed.

Customer rationality. Our analysis distinguishes between two types of customers, wary and naive. Wary customers are perfectly aware of the adviser's incentives arising both from the suitability concern ρ and from the contingent payment t that is made by provider A. To be specific, we suppose that the contract between the adviser and the product provider A is not disclosed to the customer (cf., however, the discussion of policy implications below). A wary customer, nevertheless, forms rational beliefs.

On the other hand, naive customers mistakenly believe that the quality of advice is not affected by the presence and the size of payments made by product providers. As

of a long-term contract.

[16] As part of their occupational licensing procedures, various US states require mortgage brokers to post a "surety bond" or to maintain a minimum net worth; see Pahl (2007). A surety bond is typically posted through a third party (known as surety), who is the first to be liable but is then compelled by regulation to seek redress from the broker.

discussed at length in the Introduction, survey evidence documents that customers often do not receive information about such contingent payments and that they hold, on average, beliefs that seem largely inconsistent with observed practice in the industry. What is more, even when customers could and should be aware of such payments, this may not be the most salient piece of information at the time of purchase, especially when the purchase takes place in a face-to-face situation.

Timeline. The game of contracting, advice, and purchasing proceeds in five periods. At time $\tau = 1$, product provider A chooses the price p. At the same time, a contract (T, t) is stipulated with the adviser. Given that initially we do not impose a sign restriction on T, it is inconsequential for our analysis how the bargaining power is distributed at this stage, even though it proves convenient to suppose that the product provider makes a take-it-or-leave-it offer (cf. also the discussion in Section 7). At $\tau = 2$, the adviser stipulates the fee f. Provided that the customer is willing to pay f, at $\tau = 3$ the adviser privately obtains additional information on the suitability of A or B, as represented by his posterior belief q. At $\tau = 4$, based on this information, the adviser recommends to the customer which option to choose. The game at this stage is one of cheap talk (cf. Crawford and Sobel, 1982). As we show below, the customer follows the adviser's recommendation in the only informative equilibrium.[17] At $\tau = 5$, the purchase decision is made, and then all payoffs are realized. Payoffs are not discounted and all players are risk neutral.

3 Providing advice

Given the realization of a posterior belief q (that product A is more suitable), at $\tau = 4$ it is optimal for the adviser to recommend product A whenever the adviser obtains a higher expected payoff when the customer purchases product A rather than B, i.e., when $t + qu_h + (1-q)u_l \geq qu_l + (1-q)u_h$. The adviser thus considers not only the monetary inducement t in case of recommending product A, but also the expected private costs of a subsequent mismatch, which are equal to $(1-q)\rho$ for A and $q\rho$ for B after substitution of $\rho = u_h - u_l$. If interior, the recommendation is characterized by a cutoff

$$q^* := \frac{1}{2} - \frac{t}{2\rho}, \tag{1}$$

[17] As is well known, any cheap talk game always admits a "babbling" equilibrium, in which no information is conveyed. We abstract from this uninformative equilibrium in which there is no role for advice.

so that the adviser strictly prefers to recommend A when $q > q^*$ and strictly prefers to recommend B when $q < q^*$. Note that the cutoff is not interior when $t \geq \rho$, in which case the adviser always recommends product A; for this case, we specify $q^* = 0$.

For a customer who chooses *not* to obtain advice it is optimal to always choose option B and, thereby, realize net utility

$$v_0 := v_l + \Delta_v(1 - q_0) > 0. \tag{2}$$

For this we use our simplification that in case option B consists of buying an alternative product B, this product is competitively provided at a price equal to its cost of zero. From $v_0 > 0$ we also have that the customer will always follow a recommendation to purchase B, given that the expected utility *conditional* on $q < q^*$ is strictly higher than v_0, for any cutoff $q^* > 0$. Instead, the customer's incentives to follow the recommendation to purchase product A depend on the prevailing price p. In equilibrium, however, the price p will be chosen accordingly, so that the customer also follows the advice to choose A.

Profits and surplus. Recall that we presently consider the case in which the relationship between product provider A and the adviser is governed by a two-part contract (T, t), where the fixed payment T is not subject to a sign restriction. It is then immediate that the choice of t, which governs the adviser's recommendation, will be set so as to maximize joint payoffs, i.e., the sum of the product provider's payoff

$$\Pi = [1 - G(q^*)]\,(p - t) - T \tag{3}$$

and of the adviser's payoff

$$\pi = f + T + u_l + \int_0^{q^*} \rho(1 - q) dG(q) + \int_{q^*}^1 [t + \rho q]\, dG(q). \tag{4}$$

Note that that adviser's payoff π comprises three different elements altogether: the direct fee f received from the customer; the payments from the product provider (T, t), where the contingent payment is only paid with probability $1 - G(q^*)$; and u_l together with ρ, which capture the suitability concern as $\rho > 0$ is only received when the product was suitable.

If q^* is interior, we can substitute from (1) to obtain the joint payoff of the adviser and product provider is

$$\begin{aligned} S &= \Pi + \pi \qquad (5) \\ &= f + u_l + [1 - G(q^*)]\, p + \rho L(q^*), \end{aligned}$$

where

$$L(q^*) = \int_0^{q^*} (1-q)dG(q) + \int_{q^*}^1 qdG(q)$$

denotes the ex ante probability of a *suitable* choice.

Adding the consumer surplus $v_l + \Delta_v L(q^*) - [1 - G(q^*)]p - f$ to firms' joint payoff S in (5), the total surplus in the market is equal to

$$\omega(q^*) = (u_l + v_l) + (\rho + \Delta_v)L(q^*). \tag{6}$$

Total surplus increases with the likelihood of suitable product choice, $L(q^*)$, which in turn is highest when advice is unbiased: $q^* = 1/2$. From (1), advice is only unbiased if $t = 0$.

4 Serving wary customers

In this section, we consider a market populated only by wary customers. This case provides the benchmark for our subsequent analysis of markets in which naive customers are also present. Recall that the strategic product provider A chooses both the price p that is charged to the customer and, at the same time, the two-part contract that is offered to the adviser, (T, t). After accepting this contract, the adviser is free to specify a fee f that customers have to pay before receiving advice and possibly purchasing a product.

Customer participation constraint. Recall that a customer who chooses not to obtain advice realizes the net utility v_0 from choosing option B, according to expression (2). Whether, given that the adviser applies a cutoff rule q^*, a customer follows the recommendation to purchase product A depends on the respective price p, as well as on the anticipated quality of the adviser's recommendation. To this end, a wary customer should form beliefs about the payment that the adviser receives, given that this payment affects the cutoff that the adviser applies. We denote these expectations by $\widehat{t}$ and $\widehat{q}$, respectively. Note that we presently stipulate that the payment t is not observable, which is why, at least off-equilibrium, the anticipated cutoff $\widehat{q}$ may well deviate from the true cutoff q^*.

Optimally, the wary customer follows a recommendation to purchase A if, given the anticipated cutoff $\widehat{q}$, the corresponding conditional payoff is higher than the one obtained from product B

$$v_l + \Delta_v \int_{\widehat{q}}^1 q \frac{dG(q)}{1 - G(\widehat{q})} - p \geq v_l + \Delta_v \int_{\widehat{q}}^1 (1-q) \frac{dG(q)}{1 - G(\widehat{q})},$$

which simplifies to the requirement that

$$p \leq \Delta_v \int_{\widehat{q}}^{1} (2q-1) \frac{dG(q)}{1-G(\widehat{q})}. \tag{7}$$

Intuitively, the price that the customer is willing to pay for product A is higher when it is less likely that product A is recommended (higher $\widehat{q}$), so that following a recommendation, it is more likely that product A is suitable and less likely that product B is suitable.

Next, a customer will optimally only be willing to pay a fee $f \geq 0$ up-front if the respective expected payoff exceeds that from not obtaining advice:

$$v_l + \Delta_v \int_{0}^{\widehat{q}} (1-q) dG(q) + \int_{\widehat{q}}^{1} [q\Delta_v - p] \, dG(q) - f \geq v_0. \tag{8}$$

Substituting for the customer's outside option v_0 from (2) and using the martingale property of beliefs, $\int_0^1 G(q) \, dq = 1 - q_0$, the ex ante participation constraint (8) becomes

$$p + \frac{f}{1-G(\widehat{q})} \leq \Delta_v \int_{\widehat{q}}^{1} (2q-1) \frac{dG(q)}{1-G(\widehat{q})}. \tag{9}$$

Given that $f \geq 0$, we thus conclude that this ex ante constraint implies the ex post constraint (7). As is intuitive, a customer who would optimally not follow the recommendation to purchase product A would clearly not be willing to pay a fee f to receive such advice. Hence, we need only consider for the customer the ex ante participation constraint (9).

Contract design. At $\tau = 2$, the adviser specifies the up-front fee that the customer will have to pay. If a positive fee $f \geq 0$ exists for which the customer's ex-ante participation constraint (9) is satisfied, the adviser will optimally set the fee at the highest possible level. Given a product price p and given expectations about the adviser's cutoff $\widehat{q}$, the binding constraint (9) then pins down a unique value for f.[18] Importantly, through f the adviser extracts all of the customer's residual surplus, compared to the option of choosing B without advice. This choice of f is anticipated by the product provider, who in $\tau = 1$, sets both the price p and the bilateral contract with the adviser (T, t). Recall that, for simplicity, we stipulate that the product provider can make a take-it-or-leave-it offer to the adviser, even though this assumption is inconsequential for the results in this baseline specification without agency frictions. Anticipating the adviser's subsequent choice of f,

[18] Note that we, thereby, stipulate that customer beliefs about commissions and thus $\widehat{q}$ are not affected by the adviser's subsequent choice of f.

which ensures that (9) binds, the product provider optimally chooses the fixed part T so as to make the adviser just indifferent between acceptance and rejection, so that $\pi = 0$. This implies immediately that the product provider's choice of p and t maximizes joint firm profits, $S = \Pi + \pi$.

Note now again that the actual choice of t is not observable to customers and, hence, does not affect their beliefs about the cutoff $\widehat{q}$. Consequently, to maximize joint profits, for a given product price p it is uniquely optimal to set $t = p$. As we set costs to zero, the adviser then fully internalizes joint profits when recommending A or B, and this outcome is in the interest of the product provider who fully extracts these profits through T. Note once more that the possibility to freely choose a fixed transfer T thus essentially allows to overcome the agency problem in the baseline model. We return to this observation in Section 7, when we impose restrictions on T under which agency frictions will then persist.

The optimal choice of t, for given p, is then reflected in *wary* customers' beliefs: $\widehat{t} = p$. That is, wary customers fully anticipate that a higher observed price will lead to higher commissions. Consequently, their rationally anticipated cutoff $\widehat{q}$ is given by

$$\widehat{q} = \frac{1}{2} - \frac{p}{2\rho}, \tag{10}$$

where, compared to expression (1) for q^*, we use the price p in lieu of the non-observed commission t. For given price p, wary customers' expectations (10) thus imply that despite the non-observability of t, they are not fooled by potentially biased advice. For firms this implies, in turn, that they can extract from wary customers exactly their net consumer surplus, namely by choosing p and consequently f so that the participation constraint (9) binds for the true cutoff $\widehat{q} = q^*$. Summing up, with wary customers the product provider can extract the total net surplus, $\omega(q^*) - v_0$, where $\omega(q^*)$ was defined in (6). This surplus is uniquely maximized by specifying the price $p = 0$, which gives rise to $t = 0$ and thus to unbiased advice: $q^* = \widehat{q} = 1/2$.

Proposition 1 *The equilibrium outcome with wary customers maximizes the total surplus of firms and customers. This outcome is achieved when the adviser obtains no commission and it leads to unbiased advice: $t = 0$ and $q^* = 1/2$. The customer pays a strictly positive fee for advice:*

$$f = \Delta_v \int_{1/2}^{1} (2q - 1) dG(q). \tag{11}$$

In equilibrium, advice remains unbiased, given that the adviser receives no distorting contingent payment: $t = 0$. This maximizes the joint surplus of firms and consumers, as in the present setting the adviser's sole task is to provide advice. In Section 9 we extend the analysis by introducing costly information acquisition. Though this gives rise to positive commissions even when customers are wary, the finding that contracts maximize total surplus still survives. What is key for this result is that firms can extract customer surplus through charging a fixed fee for advice, $f > 0$. As wary customers rationally anticipate the quality of advice, this makes it uniquely optimal for firms to structure incentives so that advice becomes most informative.

Even though the contingent payment t is not directly observed, in our present analysis firms can fully overcome any commitment problem vis-à-vis customers. The reason is the following. Wary customers anticipate that the product provider and the adviser choose their two-part contract (T, t) so that the adviser fully internalizes the impact of the recommendation on total firm profits. This is only the case when $t = p$. By setting $p = 0$, therefore, the product provider can credibly *commit* not to pay a positive commission. This is optimal for the product provider for two reasons: first, the subsequently chosen fixed fee for advice, $f > 0$, still allows the extraction of the consumer surplus and, second, the fixed fee in the agency contract, $T < 0$, allows the transfer of profits from the adviser to the product provider. In the next section we relax the first condition, while in Section 7 we relax the second condition.

5 Exploiting naive customers

Suppose now that customers are naive about the adviser's incentives, in the sense that they invariably hold the belief that $\widehat{q} = 1/2$. As we discussed above, one possibility is that naive customers do not understand how a specific product price affects the product provider's incentives to boost sales by paying commissions to the adviser. Alternatively, the fact that commissions are paid and that these affect the adviser's incentives may not be sufficiently salient to enter these customers' consideration when making the purchase decision.

Contract design. By the same reasoning as in the baseline case with wary customers, we only need to consider the ex-ante participation constraint for naive customers. This

is obtained from (9) simply by substituting $\widehat{q} = 1/2$. The adviser sets the fixed fee so that the participation constraint just binds, provided such a value $f \geq 0$ exists. Next, for given price p, our previous discussion of the internal agency problem between the product provider and the adviser still applies when customers are naive. That is, the product provider optimally sets $t = p$ so as to maximize joint profits $(\Pi + \pi)$, and, at the same time, sets T so as to extract the adviser's profits $(\pi = 0)$.

The key difference to the case with wary customers is that now $\widehat{q} = q^*$ holds only when $q^* = 1/2$, which in turn applies only when $t = p = 0$. For all higher prices the contingent payment is strictly positive, $t = p > 0$, so that naive customers' beliefs are consistently wrong. They underestimate the likelihood with which they will purchase product A when following advice, $q^* < \widehat{q} = 1/2$. We argue now how this, optimally, induces firm A to set the highest possible price p at which the participation constraint (9) just binds when, at the same time, $f = 0$.

Substituting f from the customers' binding participation (9) together with T from $\pi = 0$ for the adviser, we obtain for the product provider the profits[19]

$$\begin{aligned} \Pi &= \Delta_v \int_{1/2}^{1} (2q-1) dG(q) + [u_l + \rho L(q^*)] \\ &\quad + [1 - G(q^*)]\, p - [1 - G(1/2)]\, p. \end{aligned} \tag{12}$$

Intuitively, the first line reflects the customers' anticipated value from advice, given their belief that $\widehat{q} = 1/2$, and the adviser's payoff gross of his commission. The second line of (12) is zero when $p = t = 0$, so that customers' anticipated cutoff $\widehat{q} = 1/2$ equals the true cutoff q^*. Instead, for all $p > 0$ the difference is strictly positive, as then the anticipated likelihood with which product A is ultimately bought, $1 - G(1/2)$, is strictly smaller than the true probability, $1 - G(q^*)$, given that $q^* < 1/2$.

Suppose now that product provider A *increases* the product price. Through the optimal adjustment of $t = p$, the resulting change of the cutoff q^* maximizes joint profits and thus Π in (12). Applying the envelope theorem with respect to the change in q^* that is induced by the optimal change in $t = p$, the marginal change in profits is then

$$\frac{d\Pi}{dp} = G(1/2) - G(q^*). \tag{13}$$

For $p = t = 0$ (so that $\widehat{q} = q^* = 1/2$) this is zero, but it is strictly positive for all $p = t > 0$. Hence, the considered marginal increase in the product price and in the

[19] For a more formal derivation see the proof of Proposition 2.

commission, together with a reduction in the direct fee for advice, increases profits. The unique optimal choice will then imply that customers are charged no direct fee for advice, $f = 0$.

When naive customers observe a higher price for product A, they do not rationally anticipate that product provider A will also increase its commission to the adviser and that the adviser will then optimally adjust his recommendation strategy. In particular, a naive customer underestimates the probability of receiving a recommendation to buy the now more expensive product A. In fact, as the customer still expects that the recommendation to buy A happens only with probability, $1-G(1/2)$, the difference in purchase probabilities (i.e., the statistical error that is made) is exactly equal to the difference $G(1/2) - G(q^*)$ in expression (13). This observation is key. Profits thus strictly increase whenever the up-front payment for advice is reduced, provided that the participation constraint of the naive customer is still satisfied. This strict monotonicity holds because of the exploitation of the naive customer's beliefs, which are wrong whenever $t > 0$.

Once we substitute $f = 0$, together with $\widehat{q} = 1/2$, into the naive customers' binding ex-ante participation constraint, we obtain for the corresponding equilibrium product price

$$p = \Delta_v \int_{1/2}^{1} (2q-1) \frac{dG(q)}{1-G(1/2)}. \tag{14}$$

We have established the following result.

Proposition 2 *In equilibrium, naive customers are not charged directly for advice, so that $f = 0$. The corresponding price p of product A is given by (14), and the respective advice cutoff q^* is obtained from substituting $t = p$ into (1), provided this is still interior, while otherwise $q^* = 0$.*

Discussion. With naive customers, Proposition 2 thus offers a possible rationale for why frequently retail financial customers do not pay directly for financial advice. Firms generate higher profits when, in equilibrium, naive customers underestimate the true probability with which they will subsequently be advised to purchase the respective product. This makes it optimal to reduce the up-front fee as much as possible, while raising the price p and the commission t.

At the equilibrium price p for product A, together with $f = 0$, naive customers' *true* ex-ante expected payoff is strictly negative. This is an immediate consequence of why it

is optimal for firms to charge a high price for the product but no fee for advice. While this reduces total expected surplus, given that the likelihood of a suitable match L would be maximized when $q^* = 1/2$, it increases profits by extracting more surplus from naive customers, who are unaware of this. We return to this observation when discussing possible policy implications below.

Note finally that with naive customers, advice may become completely uninformative, $q^* = 0$. Then, the adviser always recommends product A. After substituting $t = p$, where p is given by (14), into the cutoff (1), this is the case when

$$\rho \leq \Delta_v \int_{1/2}^{1} (2q-1) \frac{dG(q)}{1-G(1/2)}. \tag{15}$$

6 Catering to a heterogeneous customer base

We now extend the analysis to consider a more general market composed of a fraction μ of wary customers and a fraction $1-\mu$ of naive customers. When meeting a customer, however, the adviser does not observe directly whether the customer is naive or wary.[20]

Contract design. We suppose first that the product provider has to design a single offer, p. For some retail financial services this may be a reasonable assumption. For instance, in a given "share class" that is targeted to retail investors, mutual funds typically entail a fixed load and management fee.

As a starting point, consider again the case without commissions ($t = 0$), where we also have $p = 0$, given that we set the cost to zero. Wary customers then have the same expectations as naive customers and have thus also the same willingness to pay up-front for advice. Consider now an increase in p. Naive customers then require that the fee is lowered by $df = dp[1 - G(1/2)]$, as they still hold the expectation that the cutoff $\widehat{q} = 1/2$ applies. Instead, wary customers rationally anticipate that the likelihood of being recommended product A is actually higher, as the seller optimally increases the commission t. As product A has become more expensive, for all $p > 0$ wary customers' anticipated payoff is thus strictly lower than that of naive customers.

From these observations, when there is a single offer, firms face the following two choices. When an offer shall be acceptable to all customers, the product provider sets

[20] Instead, the analysis in the previous Sections applies to the case in which the adviser directly observes whether the customer is naive or wary.

$p = t = 0$, implying that both naive and wary customers' beliefs are correct with $\widehat{q} = q^* = 1/2$. The adviser subsequently chooses the fixed fee f so as to satisfy their joint participation constraint (9). In other words, the offer is then identical to that characterized in Proposition 1. Alternatively, firms may offer a contract that is only acceptable to naive customers, in which case the product provider charges $p > 0$ as given in (14), followed by the adviser's choice of $f = 0$. Then, as customers' *true* expected payoff from turning to the adviser is negative, wary customers indeed abstain from receiving advice. There is an interior cutoff $0 < \mu^* < 1$ for the fraction of wary customers so that serving all customers with a single offer is optimal only if $\mu \geq \mu^*$, while only naive customers are targeted when $\mu < \mu^*$.

In principle, even when direct (first-degree) price discrimination between wary and naive customers is not possible, there may be scope for indirect (second-degree) price discrimination. In fact, note that the *menu* of the two offers, as characterized in Propositions 1 and 2, is incentive compatible. Naive customers are indifferent between choosing the offer designed for them or, instead, paying the up-front fee (11) in exchange for the option to buy product A at a lower price. Wary customers, however, strictly prefer "their" offer because they know that the expected payoff from the naive customers' contract is strictly negative.

Proposition 3 *When both naive and wary customers are in the market, the following outcome obtains:*
i) If only a single contract is feasible, when the fraction of wary customers is sufficiently large ($\mu \geq \mu^*$ for a cutoff $0 < \mu^* < 1$) the outcome is identical to the outcome resulting with only wary customers, as characterized in Proposition 1. Instead, when $\mu < \mu^*$, only naive customers receive advice, and the contract is identical to the outcome resulting with only naive customers, as characterized in Proposition 2.
ii) If indirect price discrimination is possible, the outcome is a menu of the contracts as characterized in Proposition 1 for wary customers and in Proposition 2 for naive customers.

Policy implications. When customers are wary, the first-best outcome with unbiased advice obtains (cf. Proposition 1). From Proposition 3 this outcome also prevails when there are not too many naive customers in the market and when firms cannot (price) dis-

criminate between wary and naive customers. In this case, the presence of wary customers protects naive customers from exploitation. This is, however, no longer the case either when there are sufficiently many naive customers in the market or when firms can successfully price discriminate between the two groups, according to assertion ii) in Proposition 3. Then, naive customers receive biased advice under exploitative terms, so that their true expected payoff is strictly below what they naively expect. In this case, policy intervention can strictly increase consumer surplus and welfare.

Specifically, policy makers could prohibit product providers from paying commissions or making other contingent payments to advisers. When $t = 0$ *must* hold irrespective of the prices and thus the margins that product providers earn, advisers would charge customers directly for advice, so that $f > 0$. Regardless of the composition of customers (μ), in this case the outcome from Proposition 1 obtains. Such a policy would represent a drastic change in some markets for retail financial services, in which customers are typically not asked to pay directly for advice and in which product providers commonly make contingent payments to intermediaries. However, a radical policy along these lines is currently being implemented in some jurisdictions, most notably by the UK's Financial Service Authority, as discussed in the Introduction. A more gradual policy change would impose a binding cap on contingent payments, though not requiring that $t = 0$. As is intuitive, in our baseline model such a cap would be preferable to no policy intervention, but it would be inferior to an outright ban on commissions.

Another policy option that is commonly adopted consists in mandating disclosure of conflicts of interest between intermediaries and customers. For the US mortgage market, by now dominated by third-party brokers, in November 2008 the Department of Housing and Urban Development has strengthened the requirement to disclose to homeowners the payments brokers receive for intermediated mortgage agreements. Similarly, since January 2008 the European Union's MiFID directive imposes mandatory disclosure for the sale of many financial products. In addition to informing customers about the level of commissions and other payments that intermediary agents receive, such disclosure policies may have the primary effect of making customers wary in the first place. Disclosure of a conflict of interest, would then act as an "eye opener" to previously naive customers.

Proposition 4 *In the baseline model, policy intervention is warranted when either the fraction of naive customers is sufficiently large or when firms can price discriminate be-*

tween wary and naive customers. The first-best outcome only obtains when either contingent payments to advisers are prohibited or when mandatory disclosure acts as eye opener by turning naive customers into wary. Consumer surplus and efficiency monotonically increase when a lower, more stringent cap $\bar{t} > 0$ *on commissions is imposed.*

In the baseline model, firm profits are strictly lower with wary than with naive customers. As long as we can abstract from the agency problem between the product provider and the adviser, which presently is made possible through the choice of the fixed part T, it is then reasonable to expect that no party will have incentives to educate naive customers, thereby eroding joint profits. We discuss this in more detail in the following Section 7.

In the baseline model, provided that disclosure works as an eye opener, mandatory disclosure has the same implication as the more interventionist policy of prohibiting commissions. In Section 9 we discuss how in a richer framework this equivalence may no longer hold. Observe also that in a market with only wary customers the imposition of either policy has presently no impact at all. This is so for two reasons. First, in the baseline model it is efficient to make no contingent payment, as only then the value of advice is largest (highest L). Second, even without policy intervention firms can achieve full commitment vis-à-vis customers, namely by setting a sufficiently low price ($p = 0$), which then makes it optimal to indeed pay no secret inducements ($t = 0$). We explore next how such a commitment problem arises once we impose restrictions on the contracts between the product provider and the adviser.

7 Dealing with agency frictions

In our preceding analysis, both charges paid by customers, p and f, were ultimately chosen to maximize joint profits. The specification of a fixed transfer T allowed to consider separately the question of how to split profits. We now suppose that $T \geq 0$ must hold in equilibrium. In standard contracting terminology, this may follow as the adviser has zero initial wealth and is protected by limited liability. More generally, as we discuss in more detail below, the imposition of such a constraint may be warranted when the relationship between product providers and advisers is more at arm's length and thus guided by short-term incentives.

Consider first the case with wary customers. In the absence of agency frictions, recall

from Proposition 1 that consumer surplus was extracted only through the fixed fee for advice, while $p = 0$ and thus $t = 0$. The product provider then made profits *only* through the fixed transfer received from the adviser: $T < 0$. When we now impose the constraint $T \geq 0$, this outcome is no longer feasible. Reconsidering the product provider's program, note first that, for given beliefs $\widehat{q}$, at $t = 2$ it is still optimal for the adviser to set the fee maximally so that the customer's participation constraint (9) just binds. Note also that even when $T = 0$ and when only $f = 0$ is feasible from the participation constraint (9), as p is set sufficiently high, we have $\pi \geq 0$ for the adviser.[21] Thus, when the product provider has all contracting power and can no longer extract surplus through a fixed transfer $T < 0$, the product provider optimally sets $T = 0$ and increases the price p so as to leave indeed no scope for the adviser to charge a positive fee for advice. Note from (9) that, for a given anticipated cutoff $\widehat{q}$, the respective maximum feasible product price is

$$p_m(\widehat{q}) = \Delta_v \int_{\widehat{q}}^{1} (2q-1) \frac{dG(q)}{1-G(\widehat{q})}. \tag{16}$$

Note next that when $p > 0$, the product provider may still have an incentive to push sales by paying a positive inducement $t > 0$, even though now T can no longer be lowered in exchange. Precisely, for given p, the product provider chooses t to maximize profits $(p-t)[1-G(q^*)]$. Taking the derivative with respect to t, we have

$$(p-t)g(q^*)\frac{1}{2\rho} - [1-G(q^*)]. \tag{17}$$

By stipulating that the hazard rate $g(q)/[1-G(q)]$ is strictly increasing, we ensure that this program has a unique solution for given p, denoted by $t^*(p)$. Choose now $\widehat{q} = 1/2$, which would prevail when customers anticipated that no commissions are paid. We stipulate that $t^*(p_m(1/2)) > 0$: At the highest feasible price for product A it is then, however, optimal for the product provider to pay commissions. From (17) this holds when ρ is not too large. Note also that $t^*(p)$ is strictly increasing, as paying a higher inducement to push sales is more profitable when the seller's margin is higher.

Wary customers hold rational beliefs, $\widehat{t} = t^*(p)$, and, consequently, expect a strictly lower cutoff $\widehat{q}$ when p increases. This gives rise to a unique price p and a respective commission $t = t^*(p)$, so that for the corresponding cutoff q^* it holds that $p = p_m(q^*)$ (cf.

[21] Precisely, this follows from our specification that $u_l \geq 0$, thus ensuring that we can treat in the same way the case in which option B represents an alternative, more basic product and the case in which it represents the option of not purchasing at all.

the proof of Proposition 5). That is, in equilibrium the product provider can charge only the price that is commensurable with his incentives to pay commissions and, thereby, bias the adviser's recommendation in favour of product A. Instead, naive customers always believe that $\widehat{q} = 1/2$, so that the product provider can charge $p = p_m(\widehat{q} = 1/2)$.

Proposition 5 *Consider the baseline model with the restriction that $T \geq 0$, given that the adviser is now constrained by zero wealth. Then, advice is biased both with wary customers and with naive customers, as in either case the product provider pays positive commissions, while there is no fixed fee for advice: $f = 0$. With naive customers, the product price is strictly higher, leading to higher commissions and thus to more biased advice, compared to the case with wary customers.*

By imposing the constraint $T \geq 0$, we restrict the product provider's ability to extract surplus from the adviser, and thus from the customer. A product provider who can only extract surplus by charging a higher price p then faces a commitment problem when commissions are not observable. In this case, the product provider has an incentive to pay commissions so as to steer advice and expand sales. In fact, recall that with wary customers a commitment *not* to bias advice in this way was obtained precisely by setting $p = 0$, which is now no longer optimal, given the restriction $T \geq 0$. More generally, we may interpret the presently analyzed case, where agency frictions persist, as a case of arm's length contracting.

Suppose now that disclosure of commissions was mandated. When customers are wary, the product provider would then optimally pay zero commissions, $t = 0$, which would then allow to charge the highest possible price $p = p_m(1/2)$. In the presence of agency frictions and with wary customers, mandatory disclosure of commissions then strictly benefits the product provider. We return in Section 11 to a comparison of policy implications under our baseline scenario and when $T \geq 0$ is imposed.

8 Competing

Given our focus on the structure of payments between customers, product providers, and financial advisers, our analysis abstracts from the institutional details of particular markets for retail financial services, such as investments or mortgages. Even in a particular class of financial products and services, there are large differences in the organization of the

industry across different countries. In what follows, we therefore analyze the effect of competition in a way that does not require spelling out the details of the market structure that prevails in a particular industry. We discuss below, however, which strategic effects in the competitive provision of advice our approach may thereby ignore.

Model extension. We still put at the heart of our analysis the provider of a more advanced product A *together* with an adviser. Recall that in our baseline model, contracts are then designed so as to maximize joint firm profits, given customers' participation constraint, which so far represented the outside option of choosing the basic product B without advice. We now envisage that customers may, instead, turn elsewhere for advice *and* for the purchase of an alternative advanced product. Precisely, we consider competition by two symmetric provider-cum-adviser networks, $i = 1, 2$, which compete in utility space by offering a given customer the anticipated expected utility $\widehat{u}_i$:

$$\widehat{u}_i = v_l + \Delta_v \int_0^{\widehat{q}_i} (1-q) dG(q) + \int_{\widehat{q}_i}^1 [q\Delta_v - p_i] \, dG(q) - f_i,$$

where p_i and f_i denote the respective payments for product A_i and advice, while $\widehat{q}_i$ denotes the anticipated cutoff that is used by the respective adviser. Note that this expression applies both to wary customers, in which case $\widehat{q}_i$ depends on the anticipated commission $\widehat{t}_i$, and to naive customers, who invariably use $\widehat{q}_i = 1/2$. To model competition, we stipulate for convenience a symmetric and continuously differentiable demand function $x_i = x(\widehat{u}_i, \widehat{u}_j)$ with $j \neq i$. From $\partial x / \partial \widehat{u}_i > 0$ and $\partial x / \partial \widehat{u}_j < 0$, where $x(\cdot) > 0$, demand for i increases when the respective expected utility $\widehat{u}_i$ increases, and it decreases when, instead, $\widehat{u}_j$ increases.

Firms' program. We can break up the firms' contract design problem in two steps. In the first step, firms determine the optimal way to deliver to customers a given utility $\widehat{u}$. Intuitively, depending on whether customers are naive or wary, the optimal contractual form mirrors that characterized in Proposition 1 and Proposition 2, respectively. That is, when customers are wary, profits will be earned through a fixed fee for advice, while when customers are naive, profits will be earned through a high product price. We denote, for given promised utility level $\widehat{u}$, the respective joint firm profits by $S^W(\widehat{u}) = \Pi^W(\widehat{u})$ when consumers are wary (using $\pi = 0$ as $T < 0$ is set sufficiently low by the product provider).

Likewise, we use for the case with naive consumers $S^N(\widehat{u}) = \Pi^N(\widehat{u})$. As naive customers' *true* expected payoff is strictly smaller than what they anticipate, for a *given* level $\widehat{u}$ we have that $\Pi^W(\widehat{u}) < \Pi^N(\widehat{u})$. However, under competition this makes it more attractive for firms to gain market share, which is the second step in the firms' program, to which we turn next.

In the second step, each firm i optimally chooses the respective level of promised utility $\widehat{u}_i$ so as to maximize expected profits $\Pi^\theta(\widehat{u}_i)x(\widehat{u}_i, \widehat{u}_j)$, where $\theta = W, N$. From differentiation and using symmetry, we obtain that, in equilibrium,

$$\Pi^\theta(\widehat{u})x_1(\widehat{u}, \widehat{u}) + \Pi_1^\theta(\widehat{u})x(\widehat{u}, \widehat{u}) = 0, \tag{18}$$

where we use the (partial) derivatives $\Pi_1^\theta(\widehat{u}) = d\Pi^\theta(\widehat{u})/d\widehat{u}$ and $x_1(\widehat{u}, \cdot) = \partial x(\widehat{u}, \cdot)/\partial\widehat{u}$. For brevity's sake we stipulate that the firms' program is strictly quasiconcave and that best-response functions (in terms of the offered $\widehat{u}_i$) intersect only once, giving thus rise to a unique symmetric equilibrium.

We capture the prevailing degree of competition in a standard and simple way, through the elasticity of demand. Given that firms essentially compete in promised utilities, in a symmetric equilibrium the demand elasticity is given by

$$\eta(\widehat{u}) = x_1(\widehat{u}, \widehat{u})\frac{\widehat{u}}{x(\widehat{u}, \widehat{u})},$$

so that the first-order condition (18) becomes

$$\Pi^\theta(\widehat{u}) = \frac{-\Pi_1^\theta(\widehat{u})\widehat{u}}{\eta(\widehat{u})}. \tag{19}$$

More intense competition is captured by an increase of elasticity everywhere. For convenience, when $\eta(\widehat{u}) = \eta$, then simply η increases.

Characterization. The introduction of competition yields now the following insights. Consider first, for a given firm, a comparative analysis in its customers' reservation value, $\widehat{u}$. This should increase when competition becomes more intense, i.e., when η increases (cf. also Proposition 6 below). With wary customers, this implies that also their true expected surplus increases by the same amount, while efficiency of advice is not affected: Advice is always unbiased. As long as total demand is elastic, however, the increase in $\widehat{u}$ leads to a standard reduction of "deadweight" welfare loss.[22]

[22] Expected welfare in a symmetric equilibrium is given by $2x(\widehat{u}, \widehat{u})\omega(1/2)$, where we substituted $q^* = 1/2$ into (6). With wary customers, the maximum surplus that a consumer can extract is $\widehat{u} = \omega(1/2)$, which

With naive customers, however, also the efficiency of advice increases with competition. The intuition for this is as follows. As customers' reservation value $\widehat{u}$ increases, the maximally feasible product price is reduced. (Note that with naive customers it always holds that $f_i = f = 0$.) Consequently, from $t = p$ also the commission decreases and thus, ultimately, the bias in the adviser's recommendation: $q^* < 1/2$. This has now an additional effect on naive customers' true expected payoff. As their reservation value $\widehat{u}$ increases, the difference between the true and the wrongly anticipated cutoff, $1/2 - q^*$, shrinks, which together with a reduction in the price implies that the difference between their true expected utility and their wrongly anticipated utility shrinks: They are exploited less.

A further insight comes now from a comparison of the cases with wary and with naive customers under competition. While we know that for a *given* reservation value, $\widehat{u}$, firms extract higher profits from naive customers, higher profits make firms compete more aggressively, which pushes up $\widehat{u}$ when customers are naive. However, we still find that firm profits are strictly higher when customers are naive, even under competition. This is so as the presence of naive customers effectively dampens competition through the following two channels. The first channel is that it is more costly for firms to increase customers' anticipated utility when they are naive. Precisely, note first that it costs firms exactly one unit of profits to increase wary customers' expected utility by the same amount, $\Pi_1^W(\widehat{u}) = -1$, given that this is obtained from a reduction in the fixed fee for advice. With naive customers, however, the corresponding loss in profits is strictly larger: $\Pi_1^N(\widehat{u}) < -1$. This follows immediately from our previous observation that an increase in $\widehat{u}$ reduces naive customers' exploitation, namely by reducing the difference between their naively anticipated utility and their true utility.

The second channel through which the presence of naive customers reduces competition is active when total demand is elastic. To see this, note first that for a *given* level of firm profits, Π, the corresponding customer reservation value $\widehat{u}$ is strictly larger with naive customers. When total demand is elastic, so that $x(\widehat{u}, \widehat{u})$ is strictly increasing in $\widehat{u}$, this would imply, for given Π, a strictly larger demand for both firms. But this makes it more expensive for firms to expand demand by increasing the promised utility, given that the

leaves the adviser and the product provider with zero expected payoff. Hence, the deadweight loss is $2[x(\omega(1/2), \omega(1/2)) - x(\widehat{u}, \widehat{u})]\omega(1/2)$.

resulting reduction in the price or the fee then applies to an already larger volume $x(\cdot)$. In other words, the larger demand that is realized when customers naively overstate their expected utility makes firms compete less aggressively.[23]

Proposition 6 *Suppose firms must compete for customers, as captured by the elastic demand function $x(\widehat{u}_i, \widehat{u}_j)$, where $\widehat{u}_i$ and $\widehat{u}_j$ represent customers' anticipated utility from two different offers. Then the following results hold:*
i) As competition intensifies, as captured by an increase in the elasticity of demand, this leads to higher consumer surplus. When customers are naive, more intense competition also reduces customer exploitation, by reducing the difference between their naively anticipated and their true expected utility from advice.
ii) Firm profits are still strictly higher when customers are naive, as the presence of naive customers dampens competition.

Discussion. Assertion i) brings out the double benefit that competition yields for naive customers, as it also reduces the scope for exploitation that arises from biased advice. Assertion ii) shows that even when competition prevails, firms still benefit when customers are naive. This has the following policy implication. When firms repeatedly interact with customers, the incremental profits that can be realized over time when customers remain naive may far exceed any immediate benefits that a product provider, together with advisers, could reap from educating customers and, thereby, gaining a larger share or even all of the market for a short time.[24] While from assertion i) competition benefits naive customers, it may thus not provide sufficient incentives for firms to educate customers.

As noted above, our analysis admittedly ignores various other strategic aspects that could arise under competition. In particular, in contrast to our simplified setting, one may allow various product providers to compete for a favorable recommendation by the same advisers, who may then stand in competition for customers. When the same or similar

[23] That demand is, in equilibrium, larger when naive customers overstate their utility may enhance efficiency if it compensates for the deadweight loss that arises when there is imperfect competition. This effect is, however, only present when total demand is elastic, and demand may also "overshoot" when competition is sufficiently intense.

[24] Precisely, such a strategy would erode naive customers' expectation of the utility obtained with the rival's offer, as well as the expectation of the utility obtained from the deviating firm's "old" offer. However, when the deviating firm, which educates customers, can react more quickly, namely by now delivering a promised utility more efficiently without biased advice, the firm's instantaneous profits may increase, next to its market share.

products are on offer at different financial intermediaries, customers may start sampling and comparing advice. Survey evidence suggests, however, that at least with retail investment products, customers seem to rarely shop for advice—and it could be conjectured that this applies, in particular, to customers who are naive about the underlying conflict of interest.[25]

9 Becoming informed to provide specialized advice

Endogenous information quality. So far the quality of the adviser's recommendation was dependent only on whether his advice was biased or not. The quality of his privately observed information was, instead, exogenous. For instance, we could imagine that the observable qualification of a financial adviser is subject to regulation. However, when products are highly specialized, it may take the adviser *additional* effort to become familiar with the customer's specific circumstances and needs. Likewise, the adviser may have to spend time and effort to himself understand the features of a particular product, most notably the advanced product A.

Denote the adviser's (privatively observed) effort by $e \geq 0$, which incurs costs $\kappa(e)$, where we stipulate that $\kappa(0) = 0$, $\kappa'(0) = 0$, $\kappa'(e) \geq 0$ for all e, and $\kappa(e) \to \infty$ as $e \to \infty$.[26] To model the resulting quality of the adviser's information, we exploit the binary structure of the match quality. Note first that any (additional) information that the adviser observes gives rise to some posterior belief, denoted by q, that product A provides a better match (i.e., that $\widehat{\theta} = A$). We characterize the quality of the adviser's information by the properties of the distribution of the posterior belief that is induced by e. An increase in effort affects the cumulative distribution function of the adviser's posterior belief, $G(q \mid e)$, by inducing a mean-preserving rotation of $G(q \mid e)$, around the prior belief, q_0:

$$\frac{dG(q \mid e)}{de} > 0 \text{ for } q < q_0, \ \frac{dG(q \mid e)}{de} < 0 \text{ for } q > q_0, \ \frac{dG(q \mid e)}{de} = 0 \text{ for } q = q_0. \tag{20}$$

For convenience, we also suppose that for all feasible effort levels $e \geq 0$ the distribution has full support on $q \in [0, 1]$ and that it is continuously differentiable in both q and e.

[25] For a large online-survey among European households, Chater, Inderst, and Huck (2010) find that while the overwhelming majority of recent purchasers of retail investment products report to obtain advice, a large majority of respondents consult only a single advisor, who is typically employed at their bank. Only a small fraction of respondents search actively for advice by consulting more than one source.

[26] Even when the time spent with customers was observable and contractible, it would be difficult to verify how hard the adviser tries to find out the best match.

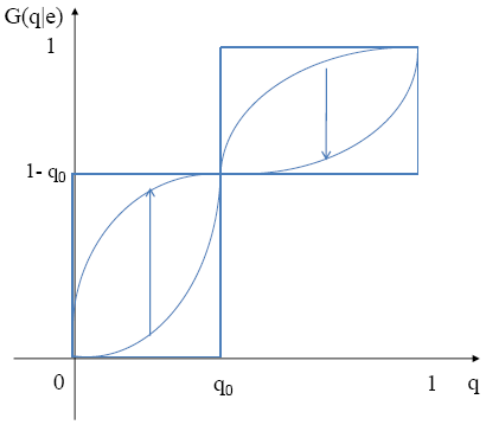

Figure 2: **Information quality.** An increase in information acquisition effort e rotates the distribution function of the adviser's belief $G(q|e)$ clockwise. The distribution is shifted upward (respectively, downward) for beliefs below (respectively, above) the prior probability q_0.

To understand the importance of condition (20), consider the extreme cases with no information and perfect information. When the adviser has access to no information, the adviser's posterior belief is always equal to the prior q_0; in this case the distribution is equal to zero for $q < q_0$ and to one for $q \geq q_0$. When the adviser has access to perfect information, the adviser's posterior belief is equal to $q = 0$ with probability $1 - q_0$ and to $q = 1$ with probability q_0; in this case the distribution is equal to $1 - q_0$ for $q < 1$ and to one for $q = 1$. As can be seen in Figure 2, the perfect information distribution is a clockwise rotation of the no-information distribution. According to condition (20), an increase in information quality results in a clockwise rotation of the distribution. Given our dichotomous structure with two states, $\widehat{\theta} = A, B$, any signal structure that results in the described rotation of the posterior distribution is more informative in the sense of Blackwell, as shown by Ganuza and Penalva's (2009) Theorem 2. This way of capturing the quality of the adviser's information is thus both intuitive and general.[27]

In what follows, we focus on the case with wary customers, for which the introduction of endogenous information quality makes a difference, in terms of both the characterization of the optimal contracts and the implications for policy. To obtain a unique solution for

[27]The distribution $G(q \mid e)$ can be generated from an underlying private signal s that the adviser observes with conditional distributions $H_A(s \mid e)$ and $H_B(s \mid e)$. See Appendix B for a characterization of the equilibrium for the specification $H_A(s \mid e) = s^{e+1}$ and $H_B(s \mid e) = 1 - (1-s)^{e+1}$ with $s \in [0,1]$ and $e \geq 0$, which satisfies the rotation ordering (20).

the choice of information quality we further assume that

$$k''(e) > \rho \max_{q\in[0,1]} \left| \frac{d^2 G(q \mid e)}{de^2} \right| \tag{21}$$

for all e, so that concavity of the maximization program is guaranteed. Without this additional assumption the equilibrium information quality need not be unique. However, standard monotone comparative statics methods can be used to extend our results also when this additional concavity assumption does not hold.

Optimal provision of effort. The adviser optimally chooses effort e to maximize the expected payoff $\pi - \kappa(e)$, where π is given in (4). When $q^* = 0$, so that the adviser always recommends A, then clearly $d\pi/de = 0$, so that the adviser has no incentive to exert effort. When, instead, $q^* > 0$ is determined by (1), expression (4) transforms to

$$\pi = [f + T + t + u_l + \rho q_0] + 2\rho \int_0^{q^*} G(q \mid e) dq, \tag{22}$$

after integrating by parts, substituting for q^*, and using $\int_0^1 G(q \mid e)dq = 1 - q_0$. Expression (22) has a simple interpretation. The first term, which is put in brackets, is equal to the expected payoff the adviser would obtain by *always* recommending option A. Note that this would allow the adviser to obtain for sure the commission t. The second term in (22) denotes the benefits, in terms of lower expected mismatch costs, when the customer makes a more informed decision based on the advice received.

When q^* is interior, then from (22) an optimal choice of effort solves the first-order condition

$$2\rho \int_0^{q^*} \frac{dG(q \mid e)}{de} dq = \kappa'(e). \tag{23}$$

For all interior q^* the left-hand side of (23) is clearly strictly positive, because the adviser cares about suitability ($\rho > 0$). The maximizing level of effort e^* is unique by our concavity assumption (21), and it is strictly positive by $\kappa'(0) = 0$. From the rotation ordering of $G(q \mid e)$ in (20), by inspecting the first-order condition (23) we immediately have the following result.

Lemma 1 *The adviser's incentives to acquire information and thus also the uniquely chosen effort e^* are hump-shaped as a function of q^* and thus also as a function of the commission t. Incentives are lowest at $q^* = 0$, which holds when $t \geq \rho$. Starting from $t = 0$*

and thus $q^* = 1/2$, as t increases also incentives increase up to $t_0 := \rho(1 - 2q_0)$, where $q^* = q_0 < 1/2$. For all higher $t > t_0$, for which $q^* < q_0$, incentives are lower.

This result is intuitive. When the adviser is *a priori* relatively sure to recommend product A, as q^* is low, the adviser has little incentive to acquire information, because this information is not likely to sway the recommendation and thus the customer's decision. At the opposite extreme, when at the prior beliefs the adviser is exactly indifferent between recommending either option, i.e., when $q^* = q_0$, any additional information will break this indifference almost surely. The adviser's incentives to acquire information are then highest.

Characterization. From Lemma 1 there are now two countervailing effects when advice becomes biased ($q^* < 1/2$) because of the payment of $t > 0$. The immediate effect is that this bias makes it less likely that the customer's choice is suitable, i.e., L decreases. The second effect is that, at least as long as still $t < t_0$, L increases as the adviser's information becomes more precise. Note now that at the unbiased recommendation cutoff, $q^* = 1/2$, the first-order effect that a reduction of the cutoff has on L is, however, zero, given that then both options are equally likely to result in a suitable choice. For all $q_0 < q^* \leq 1/2$ and thus, in particular also for $q^* = 1/2$, the effect on the adviser's quality of information is, however, strictly positive. Taken together, we conclude that L is highest when $q^* < 1/2$. Thus, advice is most informative when it is biased.

Recall now that with wary customers, we showed previously that contracts are chosen so as to maximize total surplus, i.e., the sum of firms' joint profits and consumer surplus. This insight clearly extends to the present case with endogenous information quality. For brevity of exposition, we now assume that the program to choose q^* and thus e^* so as to maximize total surplus is strictly quasiconcave.

Proposition 7 *When the adviser's information quality is endogenous, the equilibrium outcome with wary customers still maximizes the total surplus of firms and customers, which is now*

$$\omega = (u_l + v_l) + (\rho + \Delta_v)L(q^*) - \kappa(e^*). \tag{24}$$

This outcome is achieved when the adviser obtains a positive commission, $t = p > 0$, and leads to biased advice (with $q^* < 1/2$) but also to an overall higher quality of advice because

then the adviser acquires more information than would result with zero commissions and unbiased advice ($q^* = 1/2$).

Compared to the baseline model, Proposition 7 entails now the following key change in terms of policy implications. Now, when customers are wary, the imposition of a binding cap on commission or their outright prohibition interferes with efficiency. When contracts are sufficiently flexible (cf. the discussion in Section 7) and customers are wary, firms commit through the choice of prices, p, to a choice of commissions, t, that leads to the second-best outcome. Their choice maximizes total surplus under the constraint that the adviser chooses two unobservable actions: the effort e to increase information quality and the recommendation cutoff q^*.

10 Summary of policy implications

Rather that being compensated directly by customers, advisers and salespeople in the financial industry are often paid indirectly by product providers when customers decide to purchase the product offered. This practice has led to widespread claims of unsuitable advice. Policy proposals include prohibiting or, at least, seriously capping commissions, thereby also inducing intermediary agents to charge directly and more transparently for advice. However, these or other policy proposals that are meant to rectify a potential market failure can only be evaluated after having identified the precise reason for why the market does not lead to a more efficient contractual solution.

When firms face customers who are naive about the true conflict of interest that is induced by commissions, we have shown that firms can maximally exploit this naiveté by only charging customers indirectly for advice. In this case, banning commissions protects customers and, by leading to unbiased advice, increases efficiency. When customers are wary, in our baseline case without agency frictions in the supply chain we show that there is no such role for policy intervention, as firms can themselves commit to provide the highest quality of advice by setting product prices sufficiently low, thereby making it indeed not optimal to secretly increase contingent payments to steer advice. Profits are then earned (mainly) through a fixed fee for advice. In this baseline setting with wary customers and no agency frictions, hard-handed policy intervention that caps or bans commissions can easily backfire. Specifically, these policies are counterproductive in settings in which it is

necessary to pay commissions and thus to bias advice to achieve the second-best outcome, so as to increase the adviser's incentives to acquire information regarding the suitability of specialized products.

Mandatory disclosure of commissions, instead, would not interfere with firms' choice of an efficient contractual practice, even though it needs to be a sufficiently powerful "eye opener" to be effective.[28] The choice of a particular policy intervention should also depend on the perceived composition of customers in a market. When there is evidence that customers are likely to be naive about incentives, the immediate benefits of intervention are larger, and the negative side effects are smaller. As we show, intervention can then create additional efficiency gains by making it less attractive for firms to target exclusively naive customers. Also, unintended consequences of even a more interventionist policy should be a lesser concern when it is less likely that contingent payments serve an additional purpose, such as incentivizing time-consuming information acquisition in case of very specialized products.

In our baseline model, the agency problem between a product provider and the adviser can be contracted away. Technically, this is the case when the adviser is able to make a fixed transfer to the product provider. This ability to transfer resources within the supply chain can be seen as a proxy for a long-term relationship in which there are less incentives for opportunistic behavior and hence more scope to choose contractual arrangements that maximize *joint* firm profits. Instead, when fixed transfers from the adviser to the product providers are not allowed, product providers raise prices even though wary customers then rationally expect that higher inducements are paid to boost sales. Then the product provider no longer maximizes joint profits of the vertical supply chain and so does not internalize the reduction in the maximum fee that can be charged for advice. Thus, for these arm's length relationships, there is more scope for policy intervention to provide firms with a commitment device vis-à-vis wary customers.

Finally, our analysis also sheds light on the potential of competition to increase efficiency and consumer surplus. With wary customers, as can be expected, our analysis only reveals standard insights, namely that competition increases consumer surplus and,

[28] Apart from the risk of remaining ineffective, Cain, Loewenstein, and More (2005) suggest that disclosing conflicts of interest could lead to more biased advice by "morally licensing" self-interested behavior. Inderst and Ottaviani (2010a) suggest that disclosure of commissions can reduce efficiency by making sales less responsive to cost differences; however, an outright ban of commissions would be even worse.

provided that total demand is elastic, reduces deadweight loss. More interestingly, we show how competition reduces the scope for firms to exploit naive customers; by reducing prices and commissions alike, an increase in competition leads to a better alignment of the expectations held by naive customers with the actual behavior of advisers. However, eventually firms make strictly higher profits with naive than with wary customers. Thus, even in the presence of competition, the incentives for firms to educate naive customers—so as to steal market share from firms that still offer exploitative contracts—are still limited.

11 Conclusion

The present analysis is an initial step of a research program that aims at deriving positive and normative predictions on the compensation structure in the retail financial industry, with special emphasis on the role of advice. Our model allows for compensation from product providers to advising intermediaries in combination with payments made by customers to both product providers (through a price contingent on the transaction) and intermediaries (through an up-front fixed fee). Our present focus is on the role of naive vs. wary customers to explain the prevalence of different forms of compensation for advice. We also analyze how restrictions on the way product providers and advisers can resolve their internal agency problem impact on how customers pay for advice and, consequently, on the resulting efficiency of advice. Our analysis delivers a set of policy implications that tie the efficiency of different policy interventions to variables that are in principle observable, such as the customers' perception of a conflict of interest in the provision of advice or the contractual relationship between advisers and product providers.

In this spirit, future work could add more structure by analyzing the separate channels through which advisers could be disciplined, such as liability or reputational concerns. While we analyzed the potential role of competition, both in increasing efficiency and in protecting naive customers from exploitation, we also remarked that a richer model of competition could allow for additional channels for firms and customers to interact strategically. Product providers may then compete for a favorable recommendation by advisers as well as for the choice of self-directed customers. Depending on their degree of financial capability, customers may sample different advisers or rely on their own judgement. In some markets for retail financial services, product providers must also compete to be selected by product platforms (also known as wraps) to which advisers or the providers of

pension plans subscribe.[29]

Furthermore, the efficiency of making particular contractual arrangements between product providers, advisers, and customers may be impaired by various factors that remained outside our present analysis. For instance, it is often claimed that customers' up-front willingness to pay for advice is inefficiently low because they are reluctant to lock-in a certain loss. To wit, while customers pay a commission only when they decide to buy a particular product or decide to invest at all, the sure payment of an up-front fee may loom excessively large.[30] Industries may also remain stuck with a particular contractual arrangement when customers react suspiciously to any innovative offer by a maverick firm. We hope that future work will analyze the role of policy intervention for improving efficiency and protecting consumers in these circumstances.

Appendix A: Proofs

Proof of Proposition 1. Proceeding backwards from $\tau = 2$, the adviser optimally sets

$$f = f(p, \widehat{q}) = \Delta_v \int_{\widehat{q}}^{1} (2q - 1) dG(q) - [1 - G(\widehat{q})]\, p, \tag{25}$$

provided that this is feasible with $f(p, \widehat{q}) \geq 0$. Otherwise, it is not possible to satisfy customers' participation constraint (9). At time $\tau = 1$, substituting $f(p, \widehat{q})$ into the adviser's profit π, as given in (4), the product provider optimally chooses the fixed part so that

$$T = T(p, \widehat{q}, t) = -\left[f(p, \widehat{q}) + u_l + \int_0^{q^*} \rho(1 - q) dG(q) + \int_{q^*}^{1} [t + \rho q]\, dG(q) \right].$$

Once this is now substituted into (3), the product provider's profits are equal to

$$\Pi = (u_l + v_l - v_0) + \rho L(q^*) + \Delta_v L(\widehat{q}),$$

where $\widehat{q}$ depends on p according to (10). This is uniquely maximized when $\widehat{q} = q^* = 1/2$, which yields $p = t = 0$. **Q.E.D.**

Proof of Proposition 2. Proceeding as in the proof of Proposition 1, in $\tau = 2$ the adviser optimally sets $f = f(p, \widehat{q})$, which the product provider anticipates when setting

[29] For a formal analysis of regulating payments to and from such platforms, in particular under the perspective of a two-sided market, see Inderst and Valletti (2011).

[30] Chater, Inderst, and Huck (2010) report evidence from a large-scale online experiment that is at least consistent with such loss aversion.

$T = T(p, \widehat{q}, t)$. The difference is now that with naive customers, $\widehat{q} = 1/2$ remains fixed even as p and thus $t = p$ change. Thus, profits of the product provider are given by (12). The constraint $f \geq 0$ is now binding, so that expression (14) for p is obtained by substituting $f = 0$ into (9). **Q.E.D.**

Proof of Proposition 3. Take first the case of a simple offer (f, p). From the argument in the main text we have for all $p = t > 0$ that an offer that is acceptable to wary customers is strictly so to naive customers. By optimality, the participation constraint of one customer type must be binding. Taken together, this implies that the offer is characterized either by Proposition 1, when acceptable to all customers, or by Proposition 2, when acceptable only to naive customers. Denote the resulting per-customer profits by $\Pi_W < \Pi_N$. The unique cutoff for the fraction of wary customers in assertion i), μ^*, is then given by $\Pi_W = (1 - \mu^*)\Pi_N$.

For the case with a menu, note that if offering the two contracts as characterized in Propositions 1 and 2 is incentive compatible, then this is uniquely optimal. Incentive compatibility follows by construction, and strictly so for wary types, given that naive customers' true expected payoff is strictly negative. **Q.E.D.**

Proof of Proposition 4. It remains to consider the case with a cap on commissions $t \leq \bar{t}$, which can only be binding with naive customers. Then, from the argument in Proposition 2, it is uniquely optimal for firms to set $t = \bar{t}$, provided this cap binds from $\bar{t} \leq p$, with p given in (14).

The cap has no impact on the choice of $f = 0$ or p in (14). Given that the true cutoff strictly decreases with $t = \bar{t}$, while $L(q^*)$ is maximized when $q^* = 1/2$, social surplus is strictly decreasing in $\bar{t}$. Finally, naive customers' *true* expected utility is given by

$$v_0 + \Delta_v \int_{q^*}^{1/2} (2q - 1) dG(q) - p \left[G(1/2) - G(q^*) \right].$$

The derivative with respect to q^* is strictly positive from $q^* < 1/2$:

$$g(q^*) \left[p + (1 - 2q^*)\Delta_v \right] > 0.$$

Consequently, naive customers' utility is strictly increasing in q^*, when $q^* < 1/2$, and is thus strictly decreasing in the binding constraint $t \leq \bar{t}$. **Q.E.D.**

Proof of Proposition 5. Both with naive and wary customers, from the arguments in Propositions 1 and 2 it still holds at $\tau = 2$ that the adviser optimally sets $f = f(p, \widehat{q})$, as given in (25). For $\tau = 1$, recall first that even when $f = 0$ and $T = 0$, we have that $\pi \geq 0$. Together with the constraint $T \geq 0$, the product provider's profit Π is thus maximized when $T = 0$.

With wary customers, recall that the product provider optimally chooses $t = t^*(p)$ and $p = p_m(\widehat{q})$, where $\widehat{q}$ depends on the wary beliefs $\widehat{t} = t^*(p)$ (i.e., by substituting $\widehat{t} = t$ with $t = t^*(p)$ into expression (1)). As $t^*(p)$ is strictly increasing and $p_m(\widehat{q})$, with $\widehat{q} = q^*$, strictly decreasing in the true commission t, an equilibrium is unique. Existence with an interior choice $t > 0$ and an interior cutoff $0 < q^* < 1$ follows from the specification that $t^*(p_m(1/2)) > 0$ and as, from (9), we have that $p_m(q) < 0$ when q is too low. With naive customers, it is immediate that the product provider optimally chooses $p = p_m(1/2)$ and that $t = t^*(p_m(1/2))$. Finally, as $q^* < 1/2$ holds with wary customers, the respective price p is strictly lower and thus also $t = t^*(p)$ strictly lower than with naive customers. From this it follows that the respective cutoff q^* is strictly higher with wary customers. **Q.E.D.**

Proof of Proposition 6. We first derive profits $\Pi^\theta(\widehat{u})$. These are obtained from maximizing, for each pair of product provider and adviser, profits $\Pi = S$ subject to the constraint

$$v_l + \Delta_v \int_0^{\widehat{q}} (1-q) dG(q) + \int_{\widehat{q}}^1 [q\Delta_v - p]\, dG(q) - f \geq \widehat{u},$$

where $\widehat{u} \geq v_0$. As previously, we have for $\theta = N$ that always $\widehat{q} = 1/2$, while for $\theta = W$ this is obtained from the beliefs of wary customers. By applying the arguments from Propositions 1 and 2, the respective programs have a unique solution, for given $\widehat{u}$. When $\theta = W$, we have $p = 0$ and

$$f = v_l + \Delta_v L(1/2) - \widehat{u},$$

so that

$$\Pi^W(\widehat{u}) = \omega(1/2) - \widehat{u}. \tag{26}$$

When $\theta = N$, we have $f = 0$ and

$$p = \frac{v_l + \Delta_v L(1/2) - \widehat{u}}{1 - G(1/2)}.$$

This together with q^*, as obtained from substituting $t = p$ into (1), can then be substituted

to obtain firm profits with naive customers

$$\Pi^N(\widehat{u}) = u_l + \rho L(1/2) + p\left[1 - G(q^*)\right] - \widehat{u}. \tag{27}$$

With wary customers, we can now use from (19) and (26) the explicit equilibrium characterization

$$\widehat{u} = \omega(1/2)\frac{\eta}{\eta + 1}$$

to obtain $d\widehat{u}/d\eta > 0$. With naive customers, while this cannot be solved explicitly, $d\widehat{u}/d\eta > 0$ is obtained from implicit differentiation of the first-order condition (19) after substituting (27).

It remains to show that equilibrium profits are strictly higher with naive customers. This follows from inspection of the first-order condition (19), after making the following two observations. First, $\Pi_1^N(\widehat{u}) < \Pi_1^W(\widehat{u}) = -1$ holds for all $\widehat{u}$. Second, from $\Pi^N(\widehat{u}) > \Pi^W(\widehat{u})$ for all $\widehat{u}$ and strict monotonicity we have that $\widehat{u}$ is strictly higher when obtained from inverting $\Pi^N(\widehat{u}) = \Pi$ than when obtained from inverting $\Pi^W(\widehat{u}) = \Pi$, for given Π. For given Π, we can then substitute the strictly lower derivative and the strictly higher $\widehat{u}$ into the rewritten condition (19), $\Pi\eta + \widehat{u}\Pi_1^\theta = 0$. **Q.E.D.**

Proof of Proposition 7. Recall that for the case with exogenous information, an increase in the commission for selling product A results in a reduction of the cutoff q^*, and thus an an increase in the probability that product A is recommended and thus sold. We now show that this probability, $1 - G(q^* \mid e^*)$, is even higher when we take into account the adjustment of the information acquisition effort e^* that is optimally chosen by the adviser. When q^* is interior, we have for $q^* > 0$ that

$$\frac{d}{dt}\left[1 - G(q^* \mid e^*)\right] = -\frac{dq^*}{dt}\left[g(q^* \mid e^*) + \frac{dG(q^* \mid e^*)}{de^*}\frac{de^*}{dq^*}\right]. \tag{28}$$

To determine the sign of (28), recall first that $dq^*/dt < 0$ by (1). Next, from implicit differentiation of (23) we obtain

$$\frac{de^*}{dq^*} = \frac{-2\rho}{SOC}\frac{dG(q^* \mid e^*)}{de^*}, \tag{29}$$

where $SOC < 0$ denotes the second-order condition for e^*. Recall that we stipulated that the advisor's program to choose e^* yields a unique solution, which for $0 < q^* < 1$ is strictly positive. The sign of the second term in (28) is then given by $\left(\frac{dG(q^*|e^*)}{de^*}\right)^2$, which is also strictly positive. Thus, (28) is strictly positive.

From the discussion in the main text, it remains to choose q^* so as to maximize the surplus ω, as given by (24), where q^* affects e^* according to (29) and where we have to take into account the constraint $f \geq 0$. Using the binding *ex ante* participation constraint of the wary customer

$$p + \frac{f}{1 - G(q^* \mid e^*)} \leq \Delta_v \left[\int_{\widehat{q}_W}^{1} (2q - 1) \frac{dG(q \mid e^*)}{1 - G(q^* \mid e^*)} \right], \tag{30}$$

the constraint $f \geq 0$ becomes

$$\Delta_v \int_{q^*}^{1} (2q - 1) dG(q \mid e^*) - [1 - G(q^* \mid e^*)] \rho(1 - 2q^*) \geq 0. \tag{31}$$

Using the expression ω for the surplus in (24), we can also write the optimization problem with respect to the cutoff q^* as follows:

$$\frac{d\omega}{dq^*} = (\rho + \Delta_v) \left[\frac{dL}{dq^*} + \frac{de^*}{dq^*} \frac{dL}{de^*} \right] - \kappa'(e^*) \frac{de^*}{dq^*} = 0.$$

Given that e^* maximizes the adviser's payoff, so that $\rho \frac{dL}{de^*} = \kappa'(e^*)$, this becomes

$$(\rho + \Delta_v) \frac{dL}{dq^*} + \Delta_v \frac{de^*}{dq^*} \frac{dL}{de^*} = 0. \tag{32}$$

Using next, after integration by parts, that

$$\begin{aligned} \frac{dL}{dq^*} &= g(q^* \mid e^*)(1 - 2q^*), \\ \frac{dL}{de^*} &= (1 - 2q^*) \frac{dG(q^* \mid e^*)}{de^*} + 2 \int_0^{q^*} \frac{dG(q \mid e^*)}{de^*} dq, \end{aligned}$$

and substituting for $\frac{de^*}{dq^*}$ from (29), expression (32) becomes

$$\begin{aligned} \frac{d\omega}{dq^*} &= g(q^* \mid e^*)(1 - 2q^*) (\Delta_v + \rho) \\ &\quad - \Delta_v \frac{2\rho}{SOC} \frac{dG(q^* \mid e^*)}{de^*} \left[(1 - 2q^*) \frac{dG(q^* \mid e^*)}{de^*} + 2 \int_0^{q^*} \frac{dG(q \mid e^*)}{de^*} dq \right]. \end{aligned} \tag{33}$$

From (20) we have $d\omega/dq^* > 0$ when $q \leq q_0$ as well as $d\omega/dq^* < 0$ at $q^* = 1/2$. As we stipulated that the program is strictly quasiconcave, there is a unique solution $q_0 < q^* < 1/2$ and a corresponding value t from (1). However, this may not be feasible when after substituting the respective values $p = t$ into the binding constraint (30) we have $f > 0$. Then, from strict quasiconcavity the unique value q^* is the lowest value satisfying $f = 0$. Finally, $q^* < 1/2$ holds also then because at $q^* = 1/2$ we have $f > 0$, together with $t = p = 0$, so that from (30) it is indeed feasible to increase p and reduce f. **Q.E.D.**

References

Admati, A., Pfleiderer, P., 1986. A monopolistic market for information. *Journal of Economic Theory* 39, 400–438.

Bergstresser, D., Beshears, J., 2010. Who selected adjustable-rate mortgages? Evidence from the 1989-2007 Surveys of Consumer Finances, Working Paper 10-083, Harvard Business School.

Bergstresser, D., Chalmers, J.M.R., Tufano, P., 2007. Assessing the costs and benefits of brokers in the mutual fund industry. *Review of Financial Studies* 22, 4129–4156.

Bolton, P., Freixas, X., Shapiro, J., 2007. Conflicts of interest, information provision, and competition in banking. *Journal of Financial Economics* 85, 297–330.

Cain, D.M., Loewenstein, G., Moore, D.A., 2005. The dirt on coming clean: perverse effects of disclosing conflicts of interest. *Journal of Legal Studies* 34, 1–25.

Carlin, B.I., 2009. Strategic price complexity in retail financial markets. *Journal of Financial Economics* 91, 278–287.

Chater, N., Inderst, R., Huck, S., 2010. Consumer decision-making in retail investment services. Report to the European Commission (SANCO).

Chen, J., Hong, H., Kubik, J.D., 2006. Outsourcing mutual fund management: firm boundaries, incentives, and performance. Working paper, University of Southern California, Princeton, and Syracuse.

CFA Institute, 2009. *European Union Member Poll on Retail Investment Products*. Summary Report.

Commission of the European Communities, 2009. *On the Follow Up in Retail Financial Services to the Consumer Markets Scoreboard*. Commission Staff Working Document 1251.

Crawford, V.P., Sobel, J., 1982. Strategic information transmission. *Econometrica* 50, 1431–1451.

Cummins, D., Doherty, N., 2006. The economics of insurance intermediaries. *Journal of Risk and Insurance* 73, 359–339.

DellaVigna, S., Malmendier, U., 2004. Contract design and self-control: theory and evidence. *Quarterly Journal of Economics* 119, 353–402.

Federal Trade Commission, 2008. *Before the Board of Governors of the Federal Reserve System: In the Matter of Request for Comments on Truth in Lending.* Proposed Rule Docket No. R-1305.

Financial Services Authority, 2009. *Distribution of Retail Investments: Delivering the RDR.* Consultation Paper 09/18.

Financial Services Authority, 2010. *Quality of Advice on Pension Switching: An Update.* Report April 2010.

Ganuza, J.-J., Penalva, J.S., 2010. Signal orderings based on dispersion and the supply of private information in auctions. *Econometrica* 78, 1007–1030.

Gravelle, H., 1993. Product price and advice quality: implications of the commission system in life assurance. *Geneva Papers on Risk and Insurance Theory* 18, 31–53.

Gravelle, H., 1994. Remunerating information providers: commissions versus fees in life insurance. *Journal of Risk and Insurance* 61, 425–457.

Hackethal, A., Inderst, R., Meyer, S., 2010. Trading on advice. Unpublished working paper, University of Frankfurt.

Inderst, R., Ottaviani, M., 2009. Misselling through agents. *American Economic Review* 99, 883–908.

Inderst, R., Ottaviani, M., 2010a. Competition through commissions and kickbacks. *American Economic Review* forthcoming.

Inderst, R., Ottaviani, M., 2010b. Sales talk, cancellation terms, and the role of consumer protection. Unpublished working paper, Northwestern University and University of Frankfurt.

Inderst, R., Valletti, T., 2011. Regulating platform charges. FSA Working Paper.

Jackson, H.E., Burlingame, L., 2007. Kickbacks and compensation: the case of yield spread premiums. *Stanford Journal of Law, Business & Finance* 12, 289–361.

Keith, E., Bocian, D., Li, W., 2008. Steered wrong: brokers, borrowers, and subprime loans. Working paper, Center for Responsible Lending.

Malmendier, U., Shanthikumar, D., 2007. Are small investors naive about incentives? *Journal of Financial Economics* 85, 457–489.

Pahl, C., 2007. A compilation of state mortgage broker laws and regulations, 1996–2006. Federal Reserve Bank of Minneapolis, Community Affairs Report No. 2007–2.

Stoughton, N.M., Wu, Y., Zechner, J., 2011. Intermediated investment management. *Journal of Finance* forthcoming.

US Department of Treasury, 2009. *Financial Regulatory Reform. A New Foundation: Rebuilding Financial Supervision and Regulation.* White Paper.

Appendix B: Example

To illustrate Section 9's model with endogenous information acquisition and to obtain some additional comparative statics results we turn to a simple parametric example. This example also allows us to show how the distribution of the adviser's posterior beliefs, $G(q \mid e)$, can be derived from a noisy signal technology.

Suppose that the adviser privately observes a signal $s \in [0,1]$ with conditional distributions $H_A(s \mid e) = s^{e+1}$ and $H_B(s \mid e) = 1-(1-s)^{e+1}$ parametrized by $e \geq 0$. The adviser's posterior belief as a function of the observed signal is then equal to

$$q = \widetilde{q}(s) := \frac{q_0 s^e}{q_0 s^e + (1-q_0)(1-s)^e}$$

by Bayes' rule. Note that $\widetilde{q}(0) = 0$ and $\widetilde{q}(1) = 1$. Also, we may now, alternatively to the specification of a cutoff q^*, define a cutoff on the signal s^* with

$$\frac{q_0}{1-q_0}\frac{1-q^*}{q^*} = \left(\frac{1-s^*}{s^*}\right)^e,$$

so that the adviser recommends A if $s \geq s^*$ and B if $s < s^*$. After some transformations, the likelihood of a suitable choice as a function of s^* is then given by

$$L = 1-(1-q_0)(1-s^*)^{e+1} - q_0(s^*)^{e+1}.$$

Given that the signal has the unconditional cumulative distribution function $q_0 H_A(s \mid e) + (1-q_0)H_B(s \mid e)$, we further obtain

$$G(q \mid e) = q_0\left[\widetilde{q}^{-1}(q)\right]^{e+1} + (1-q_0)\left[1-\left[1-\widetilde{q}^{-1}(q)\right]^{e+1}\right]. \tag{34}$$

It is straightforward to show that this $G(q \mid e)$ satisfies the rotation ordering (20).

For a comparative analysis we specify that the information acquisition cost is quadratic, $\kappa(e) = e^2/(2c)$ with $c > 0$. With this specification, we now analyze how the outcome depends on the likelihood with which the advanced product A is ex ante more suitable, q_0. For Figure 3 we specify $\rho = 0.75$ for the adviser's preferences, $\Delta_v = v_h - v_l = 2$ for the incremental benefits of a suitable choice, and $c = 0.65$ for the adviser's cost of effort function. As q_0 decreases, the basic option (or, equivalently, the option of not buying) is ex ante more likely to be suitable; alternatively, product A is targeted more to a niche market. As illustrated in the figure, under the optimal contractual arrangement with wary customers, the commission t paid to advisers increases and the recommendation cutoff q^* decreases when the initial probability q_0 is reduced from 1/2 to 0. While a

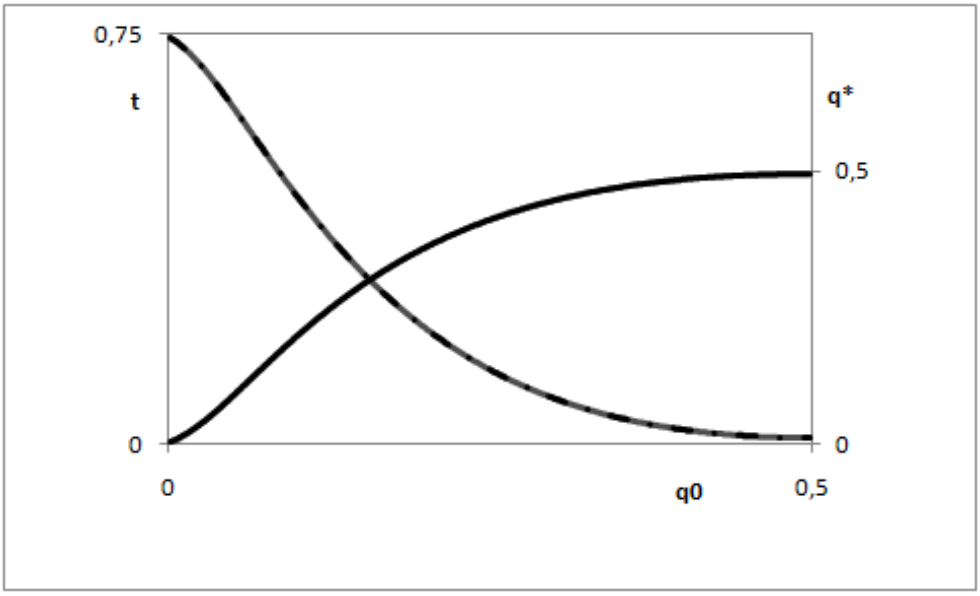

Figure 3: **Commissions and bias.** For the parametric example discussed in the text, this figure reports the equilibrium level of commission t (the decreasing curve) and the equilibrium recommendation cutoff (the increasing curve) as a function of the initial probability q_0 that product A is suitable.

recommendation becomes thus more and more biased, in this example the loss in the quality of advice generated by the bias is more than compensated by the higher level of information acquisition that is thereby induced.