Quality Disclosure and Regulation: Scoring Design in Medicare Advantage [*]

Benjamin Vatter[†]

October 3, 2021

Job Market Paper

Click here for the latest version and **here for the appendix**

Abstract

Informing consumers of the quality of products alters their demand and, therefore, firms' incentives to invest in quality. By leveraging this mechanism, regulators could coordinate quality disclosure and regulation policies, reduce spending, and improve welfare. I study how to design policies that inform consumers and regulate quality using information alone. Combining data and theory, I examine the design of one of the most extensive disclosure programs in the United States: the Medicare Advantage Star Ratings. Quality heterogeneity in this health insurance market is considerable, costs billions in subsidies, affects population health, and is difficult to assess without additional disclosure. I specify and solve an optimal scoring design problem for this market and find an alternative welfare-improving design. The new scores use the same data as the Star Ratings, are less complex, shift demand towards higher quality, induce higher investments by firms, reduce government spending, and address a common multitasking problem. The analysis provides insights into why quality certifications are effective, what governs the coarseness of disclosure policies, and why informational campaigns accompany successful disclosure policies. Overall, my alternative design increases total welfare by $650 per Medicare beneficiary year, with half of the gains stemming from better information and half from higher induced quality.

[*] I would like to thank David Dranove, Igal Hendel, Gaston Illanes, and Amanda Starc for their invaluable mentorship and advice. I thank Vivek Bhattacharya, Mar Reguant, Robert Porter, William Rogerson, Molly Schnell, Sebastian Fleitas, Jose Ignacio Cuesta, Carlos Noton, Victoria Marone, Matthew Leisten, and seminar participants at Northwestern University and the NHEBA work group, for their valuable comments and suggestions. This work benefited from generous funding by the Robert Eisner Graduate Fellowship. All errors are my own.

[†] Northwestern University. Email benjaminvatterj@u.northwestern.edu

1 Introduction

Quality scores are ubiquitous. They grade the quality of schools, insurance products, energy efficiency in construction and appliances, and a growing list of products and services. Scores help consumers choose when information is scarce and, by doing so, alter firms' incentives to invest in quality. Through these effects, scores meaningfully impact consumer surplus (Dranove and Jin, 2010), and the quality of products firms offer (Barahona et al., 2020). Governments, who play a prominent role as certifiers of quality and designers of scores, are interested in leveraging these supply and demand effects for the market's betterment. The challenge is that how to design scores that improve welfare is, with few exceptions (Zapechelnyuk, 2020), unknown.

This paper investigates scoring design empirically by studying the Medicare Advantage (MA) health insurance market. Scores in MA summarize many quality dimensions, including disease management protocols and the quality of providers in each plan's network. These scores' design varies yearly, data on the market is readily available, and quality heterogeneity impacts mortality (Abaluck et al., 2021) and public spending (CMS, 2016). These features make MA well-suited for the study of scoring design and a setting where improvements might have significant welfare consequences. This paper evaluates the potential impact of redesigned scores by developing and estimating a model of the market and solving the designer's problem using a novel method that draws on insights from the theory on information design (Kamenica and Gentzkow, 2011). The results indicate that relative to a full-information social optimum, insurers underinvest in quality and substantial surplus is lost due to consumers' limited information. Moreover, there is pervasive misclassification wherein consumers would prefer, ex-post, some lower-scoring products to higher-scoring ones. The new design reveals mechanisms through which scores can coordinate consumers to demand more quality from insurers, improve consumers' information about available products, and reduce misclassification. The exercise shows how scoring design might significantly improve welfare in various markets with an endogenous provision of quality and limited consumer information.

Designing welfare improving scores is challenging for several reasons. First, if insurers respond to scores by adjusting quality, then it is no longer the case that a more informed consumer is better off. For example, a system that awards certifications to high-quality plans provides less information than one with an additional level for medium-quality. However, replacing the first design with the second might lead some insurers to reduce their investments and decrease welfare. The market fails to correct this loss as uninformed consumers do not generate the appropriate incentives for firms to invest in quality.[1] Therefore, the designer has to make a judicious choice of how much information

[1] For example, in markets with a unique efficient level of quality production, dichotomous certification can lead to efficient investments (see Section 2). A three level score, instead, might lead firms to differentiate in quality to reduce

to share with consumers. To make this decision, they have to learn the social value of quality, its production cost, and how changes to quality and information affect prices and competition. This is the designer's second challenge. Finally, a third challenge is using this information to design a new system. The designer can not solve the associated functional optimization problem using standard techniques or unguided exploration, as scores are discontinuous functions and the space of alternatives is prohibitively large. Moreover, evaluating any guess requires computing a non-trivial counterfactual equilibrium of investments, prices, demand, and beliefs.

I begin by documenting that consumers and insurers respond to changes in the rating design. I contribute to extensive evidence about demand responses to MA ratings (Dafny and Dranove, 2008; Reid et al., 2013; Darden and McCarthy, 2015) by documenting that consumers respond in level and composition. This finding states that, for example, consumers prefer plans with four stars to those with three and, all else equal, value the difference between these ratings more when the system demands larger reductions in hospital readmissions to achieve this increase. Additionally, I show that quality in MA responds to rating incentives as it does in other markets (Jin and Leslie, 2003). Using variation produced by the introduction of new quality metrics to the Star Rating formula, I develop a triple-difference strategy that shows that plans at risk of losing ratings increase their quality. Together, the demand and supply responses validate the necessary conditions for quality-regulating disclosure. Leveraging this variation in scores, demand, and supply, I develop and estimate a structural model of the market.

The model builds upon the demand and price competition models of Curto et al. (2021a), and Miller et al. (2019), extending them in two ways. First, I incorporate an investment stage in which insurers invest in multiple quality dimensions. For example, an insurer can reduce readmission rates by adding better hospitals to its network or can increase flu-vaccination rates by contracting with more pharmacies. Second, I model consumers as maximizing an expected utility subject to uncertainty over quality. Thus, the model gives a specific interpretation to the descriptive findings: Consumers prefer four over three stars because they understand that plan quality must be higher to obtain a higher rating. They value the difference between ratings more in years in which the system better reflects improvements in their most valued quality dimensions. These two extensions to the model allow me to simulate equilibria under counterfactual scoring designs.

I use the estimated model to find new scoring designs that maximize total welfare. I focus on designs that deterministically assign higher scores to higher quality, using only a finite amount of scores. This class of finite monotone scores incorporates many standard disclosure policies, such as the MA Star Rating system, car-safety certifications, restaurant hygiene grades, food labels,

price competition, thus decreasing welfare, as in Ronnen (1991).

and more.[2] To tackle the computational challenge involved in finding optimal scores, I show that they can be expressed as a composition of two easy to optimize functions. To address the cost of solving many counterfactual equilibria during optimization, I leverage the intuition of Kamenica and Gentzkow (2011) that choosing a scoring system is, to some extent, equivalent to choosing a distribution over consumers' posterior beliefs. I use this to develop a computational approach that first solves a large set of counterfactual scenarios and then identifies the value of every alternative scoring design as a distribution over these values. The reformulation and computational approach render the scoring design problem manageable.

The main results of this paper explore alternative scoring rules that improve total welfare, revealing broad lessons for scoring design. The results suggest that the optimal disclosure policy when firms provide an inefficient level of quality often involves limiting consumers' quality information. In MA, I find that insurers under-invest in quality relative to the socially efficient level both in the status quo and under a counterfactual of full information. Coarsening consumers' information by pooling intervals of low quality while providing accurate signals of high quality makes products obtaining low scores less attractive to consumers, jolting investment. Moreover, the results indicate that total welfare is not monotonic in how informative a score is ex-ante. Coarser scores can lead to lower quality heterogeneity, less uncertainty for consumers, and higher quality overall.

The main alternative design proposed in this work improves total welfare by $650 per consumer-year while relying on the same inputs and technology as the Star Ratings. This design improves the status-quo by leveraging the regulatory power of disclosure on quality production in two ways. First, the new design provides consumers with coarse information about low and medium qualities but a precise signal for high quality. This structure shifts demand towards high-quality products, making it profitable for firms to increase their quality, thus improving welfare.[3] Second, the new design addresses a fundamental multitasking problem (Holmstrom and Milgrom, 1991). Intuitively, every scoring system establishes different paths for plans to achieve the same rating. For example, a plan can reach four stars by having all of the best outpatient clinics in its network and a few good inpatient hospitals or all the best hospitals and a single outpatient clinic. The plan will pick the cheapest path, while consumers and the regulator might prefer a different one. The new design improves the alignment between firms' incentives to substitute investments across quality dimensions with consumers' preferences. Thus, the new system increases both quality investment, and how that investment is allocated across different dimensions of quality. Consumers

[2] Dworczak and Martini (2019) prove that similarly defined scores can be optimal in a wide array of scenarios, albeit with exogenous quality. Their definition allows for specific segments of the score to be fully revealing. I explore these designs in the appendix.

[3] Harbaugh and Rasmusen (2018) show that coarse information is optimal due to similar logic in a theoretical framework with exogenous quality and voluntary participation.

lose information about lower qualities, but as quality is higher and less heterogeneous, they make fewer ex-post mistakes when choosing insurance. Overall, consumer surplus increases by $131 per consumer-year or 2.1 months of average market premiums. A decomposition of welfare gains indicates that about half of the improvements in surplus stem from better information and the remainder from the endogenous change in quality.

Exploring how the solution balances regulatory power and disclosure delivers additional insights for scoring design. There are two extremes to this balance. On one end, a fully informative design forgoes any attempt to regulate and instead maximizes disclosure.[4] On the other end lies quality certification (i.e., a system with two scores). This common class of designs, often shown as a sticker or label, pools all qualities below a threshold together and analogously above it.[5] This pooling creates the harshest penalty for firms that fail to meet the desired goal of the regulator (the threshold quality) and the largest gain for those who do. Certification delivers minimal information to consumers, yet if firms react by delivering precisely the threshold quality, consumers might be as well informed about their available options as they were before. I confirm that the gains in quality can offset losses in information, showing that the optimal certification design for MA attains 97% of the welfare gains of the more sophisticated solution.

Certification also sacrifices valuable product heterogeneity, as some firms find it too costly to exceed the threshold quality. Adding additional cutoffs below the certification threshold allows these firms to participate, but in turn, diminishes the incentive for firms to provide high quality. This trade-off limits the number of different scores used by an optimal design, explaining why disclosure systems are often coarse. The coarseness of these scores induces a discontinuity in insurer revenue as quality increases, known as the ”cliff effect.” These cliffs effects are fundamental in providing incentives for firms to increase their quality. The results of this paper suggest that doing away with cliff effects, as recently suggested to policymakers (MedPAC, 2020), would likely result in lower quality.

The analysis also reveals two reasons why public information campaigns are invaluable for centralized scoring designs. First, I show that scores are ineffective in steering consumers towards products they dislike. Intuitively, they are less likely to find helpful information in a system that systematically embellishes products they dislike. The implication for MA is that the regulator’s attempt to nudge patients towards products that improve long-run health outcomes (e.g., by managing the populations’ diabetes) erodes the regulatory and informational value of the scoring

[4] Without additional structure, fully informative designs are infeasible when aggregating multiple continuous dimensions into a single one. If consumers aggregate information into a common index, and this index is known to the designer, then full information is feasible.

[5] Common examples of certification are the USDA or EU organic labels, front-of-package warning labels for sugars and calories, Energy Star certification on computers and monitors, NAHQ certification of medical professional quality, and a plethora of ISO certifications ranging from upper management to food processing technology quality.

system. However, if she could educate the population about the value of her preferred plans and change their preferences, these frictions would not exist. This finding might explain the role of the public information campaign that accompanied the successful Chilean food labeling program (Reyes et al., 2019). Second, I prove that consumers' preferences for quality are only identified from their plan choices if they understand the scoring design. Thus, lacking additional preference data, the regulator can not determine the optimality of its scoring design. This observation, however, does not preclude rating systems from improving welfare or increasing product quality. For example, I show that in MA, reassigning the existing ratings under a different rule would improve welfare and increase quality, even if consumers are unaware of the change.

This work contributes to a growing literature studying the design of disclosure policies. By solving the optimal design within a broad class of scores in an empirical setting, I bridge the gap between the theoretical literature that searches for optimal designs (Albano and Lizzeri, 2001; Rodina and Farragut, 2016; Boleslavsky and Kim, 2018; Ball, 2019; Zapechelnyuk, 2020) and the empirical literature that measures the impact of existing designs (Bollinger et al., 2011; Werner et al., 2012; Elfenbein et al., 2015; Chen, 2018; Araya et al., 2018; Houde, 2018b).[6] The analysis shows that the variation required to study the impact of disclosure is similar but less demanding to the one needed to solve the design problem. It also delivers methods to solve the latter. Given the focus on quality responses, this work speaks to research on the supply-side effects of centralized mandatory disclosure, studied in education (Mizala and Urquiola, 2013; Allende et al., 2019), health care (Chou et al., 2014), airlines (Forbes et al., 2015), and electrical appliances (Houde, 2018a), to name a few. I contribute to this extensive literature by showing how regulators might leverage quality responses as a regulatory tool. This work also reveals that some of the flaws documented in nursing homes scores (Feng Lu, 2012), energy-efficiency certifications (Clay et al., 2021), and schools scores (Neal and Schanzenbach, 2010) are expressions of the same multitasking design flaw, which share a common solution.

This study also adds to extensive literature on quality provision and regulation. Concerns about the inefficient provision of quality are foundational (Arrow, 1963), and their origin in imperfect competition has been explained theoretically (Spence, 1975; Mussa and Rosen, 1978; Schmalensee, 1979), and documented empirically (McManus, 2007; Crawford et al., 2019). I document that quality is underprovided in MA and would be so even under perfect information. This paper's solution to quality degradation through disclosure results in an outcome similar to licensing, with low-quality products receiving negligible demand. Exclusionary licensing has been studied intensively in education (Angrist and Guryan, 2008; Larsen, 2014; Larsen et al., 2020), occupational licensing (Barrios, 2017; Kleiner and Soltas, 2019; Farronato et al., 2020), and pharmaceuticals (Atal et al., 2021). A common negative finding in this literature is that prices increase without changing

[6] See Dranove and Jin (2010) for a review of earlier work on quality disclosure.

quality. My results indicate that this is unlikely to happen with quality scores as failure to increase quality would result in a less differentiated product market, intensifying price competition.[7] The analysis behind this result, and others in the paper, builds on an extensive literature that studies competition with endogenous non-price product attributes (Berry and Waldfogel, 2001; Gandhi et al., 2008; Nosko, 2014; Fan, 2013; Berry et al., 2016; Hui et al., 2018; Fan and Yang, 2020). I contribute to it by studying competition with multidimensional investments with risky outcomes.

Finally, this work provides a policy contribution. The alternative design I find is implementable using the same data and technology as the Star Rating. The results do not require the regulator to alter its parallel efforts to regulate quality or prices. It is also deterministic, which allows the regulator to communicate clear goals to the industry and make the design transparent to consumers. This work also contributes to a broad literature studying the industrial organization of MA using structural methods (Town and Liu, 2003; Lustig, 2010; Aizawa and Kim, 2018; Curto et al., 2021a; Nosal, 2011; So, 2019; Charbi, 2020). Particularly related is Miller et al. (2019) who study optimal subsidy design in MA, taking quality provision and regulation as given. In contrast, I consider subsidies and many non-price non-quality attributes (e.g., deductibles) studied by Miller et al. as exogenous. Both works are, therefore, complementary.

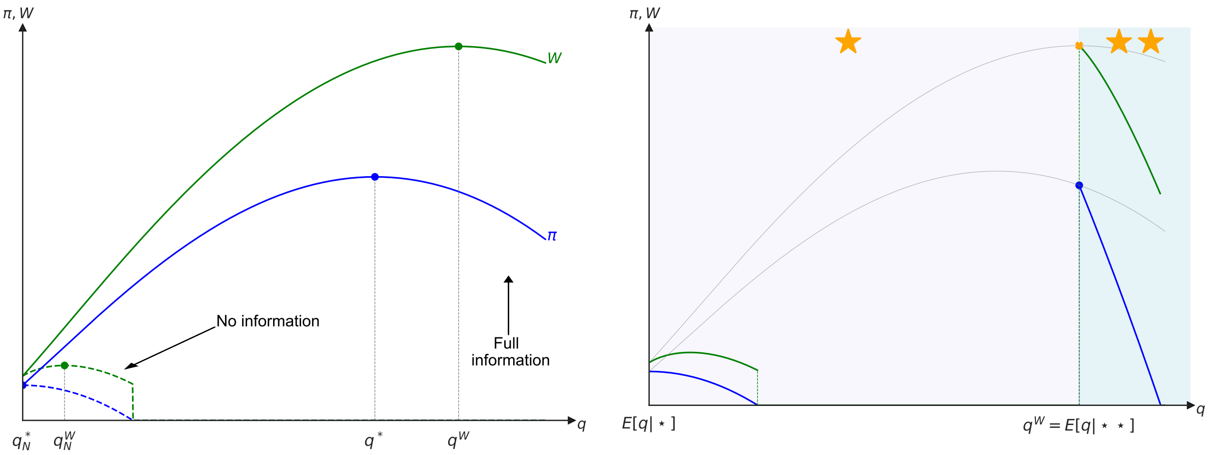
I organize the remainder of the paper as follows. Section 2 illustrates the regulatory role of disclosure policy using the single-product monopolist model of Spence (1975). Section 3 presents the setting and describes the MA Star Rating system. Section 4 studies the effects of this rating on demand for insurance and the supply of quality. These findings clarify the potential for MA ratings to regulate quality and the variation leveraged in the structural model. Section 5 presents the structural model. Section 6 describes how I estimate the model and how the data variation identifies it. Section 7 develops the main analysis showing the alternative scoring design and how variations affect welfare in the market. Section 8 discusses the robustness of the findings to alternative assumptions about consumers' understanding of the rating design. Section 9 concludes.

2 Quality-Regulating Disclosure

I begin by describing the economic intuition underlying disclosure's ability to inform consumers and regulate quality simultaneously.[8] Following Spence (1975), consider a monopolist that chooses a quality q and price P for his single indivisible product. Each consumer decides whether to buy a

[7] Additionally, scores can be easily reassigned under an unexpected failure. For example, after receiving criticisms for its hospital rating design in 2016, CMS rapidly corrected erroneous assignments and changed its system the following year.

[8] In order to focus on the primary forces used by the designer, I abstract away from unobserved investments, multidimensional quality, and other regulatory concerns found in the application.



(a) Informational extrema

(b) Regulated

Figure 1: Quality certification under monopolistic provision

Notes: This figure illustrates how quality certification changes a monopolist's incentive to invest in quality. On the left, the profit and welfare curves for full and no information scenarios are presented. In the right figure, π and W are distorted due to the effect of $\psi(\cdot)$ on consumer beliefs, and vanish when the participation constraint of the monopolist is violated. The gray lines trace the former full-information curves. The shaded areas represent the distinct rating segments.

unit of the good, resulting in an inverse demand $P(x, q)$, where x denotes quantity. The monopolist has a production cost of $C(x, q)$ resulting in profits of $\pi(x, q) = xP(x, q) - C(x, q)$. The market's total welfare is

$$W(x, q) = \int_0^x P(v, q)dv - C(x, q)$$

For any quantity x, the profit-maximizing monopolist chooses a quality that equates the marginal revenue of quality to its marginal cost ($xP_q(x, q) = C_q(x, q)$). In contrast, total welfare is maximized at the point where the marginal cost of quality equals the marginal consumer surplus ($\int_0^x P_q(v, q)dv = C_q(x, q)$). As marginal revenue and marginal surplus typically differ, the monopolist's optimal choice of quality will, in general, be inefficient.

Figure 1.a illustrates this observation. The solid lines labeled as "full information" show a scenario in which the monopolist degrades quality by choosing q^* instead of q^W. When a firm maximizes its profit under market pressures at an inefficient level of quality it is said that it has market power over quality (Crawford et al., 2019). The wedge between the efficient (q^W) and the profit-maximizing choices (q^*) is known as the *Spencian* distortion. The presence of such distortions is often used to justify the need for quality regulation. For example, the Spencian distortion could be eliminated using a minimum quality standard that requires products in the market to have at least quality q^W.

Now suppose that consumers can not ascertain the product's quality before purchase.[9] Consumers have rational expectations and thus, lacking any quality information, assume that the monopolist would provide his cost-minimizing quality.[10] The ”No information” dashed curves in Figure 1.a illustrate this outcome for a strictly increasing cost function. The monopolist has no incentives to increase quality under this scenario, and substantial welfare is lost relative to full information.

A regulator observes this failure and intervenes by measuring the product's quality and disclosing a signal to consumers. Consumers and the monopolist know the structure of this signal and respond accordingly. For example, the regulator can recover the full information scenario by making her disclosure signal equal to the the measured quality. This choice would eliminate the information asymmetry frictions but restore the Spencian distortion. Alternatively, the regulator might recognize that, conditional on quality, her disclosure rule changes the firm's demand but not the product's value to consumers. Therefore, disclosure affects the firm's objective but not the (ex-post) welfare function, providing an opportunity to align the two objectives. Figure 1.b illustrates the outcome of a policy that shows consumers one star if quality falls below the socially efficient level q^W, and two stars if it is equal or exceeds it. This design is a quality certification, typically shown with a label on the product (e.g., USDA Organic or Energy Star). As illustrated, this disclosure policy disrupts the firm's profit curve. Choosing to deliver any one-star quality lands the monopolist in the scenario of ”No information,” as consumers again believe quality to be minimal. Investing in two stars sets the monopolist in a new scenario in which consumers believe quality to be precisely q^W. In this case, the profit and total welfare curves decrease and begin at the full-information efficient-quality point. The monopolist maximizes profit by choosing the efficient level, and the Spencian distortion disappears. Consumers purchase a product of efficient quality with accurate beliefs.

This illustration reveals that the regulatory power of scores stems from their ability to marshal demand. By revealing a coarse signal, the regulator coordinates consumers to demand high quality, offsetting the monopolist's market power. I formalize this intuition in Appendix 2.1, showing that under standard assumptions, quality disclosure leads to weakly greater welfare than full information because it partially or fully eliminates the Spencian distortion. Additionally, in Appendix 2.2 I follow Zapechelnyuk (2020) and derive the optimal disclosure rule for a scenario in which the monopolist has private information about his investment cost. Overall, the regulator must limit consumers' information to regulate quality, confronting her with a trade-off between

[9] I assume that there are not alternative mechanisms to achieve the full-information outcome. For example, the monopolist can not credibly commit to producing a certain quality, or the market is short-lived.

[10] No other belief can be held in a rational expectations equilibrium as the monopolist always has incentives to reduce quality to save on costs. As consumers can not observe this deviation, the demand for the product remains the same, hence increasing the monopolist's profits.

her disclosure and regulatory goals. To determine the right balance between the two, she must understand the cost of quality and consumers' preferences. The same components will have to be understood and leveraged to form a new scoring design for the empirical application that follows. However, the theoretical solution approach I use in the appendix to derive optimal designs does not lend itself to an imperfectly competitive setting, thus requiring a computational replacement.

3 Institutional Details and Data

3.1 Medicare Advantage and The Star Rating Program

Since 1965, retirees and disabled individuals in the US have had access to a public health insurance system known as Medicare. This system provides hospital, physician, and outpatient coverage under a publicly administrated and highly subsidized scheme. A series of reforms enacted between 1982 and 2003 established an alternative to traditional Medicare (TM), known today as Medicare Advantage (MA). Under MA, the Center For Medicare and Medicaid Services (CMS) contracts with private insurance companies to provide alternative coverage for Medicare beneficiaries in exchange for a prospective risk-adjusted capitated payment. Enrollment in MA has been steadily increasing during the past decade; out of the nearly 65 million Medicare-eligible enrollees in 2019, 34% chose a plan in MA.[11]

MA markets are highly concentrated and regulated. During 2019, the average market (county) had 90% of its enrollment controlled by only two firms. At the national level, four firms command 69% of all enrollment (Frank and McGuire, 2019). In most counties, insurers offer a wide array of plans differing in their coverage generosity (e.g., coinsurance, deductibles) and their access to clinical quality. CMS strictly regulates the financial characteristics of plans, including minimum requirements on coverage generosity and limits on premiums relative to coverage. Curto et al. (2021b) provide further description of price and coverage regulation, which I complement with extensive detail in Appendix Section 3.1. CMS also subsidizes consumers by paying a large fraction of plan premiums. Nearly half of all MA plans are offered at a zero premium to consumers.[12]

In contrast to financial characteristics, differences in plan quality are less regulated and harder for consumers to ascertain. These differences are due to variation across plans in the size and makeup of provider networks, disease management protocols, and processes for approving costly

[11] Traditional Medicare is composed of part A (hospital coverage) and part B (physician and outpatient coverage). For further details on the history of this program, see McGuire et al. (2011). For details on risk-adjustment and residual selection, see Brown et al. (2014) and So (2019).

[12] MA consumers still pay their part B premiums. However, this amount is paid regardless of their choice of TM or MA.

medical procedures, among other factors. Information regarding these aspects of plans are rarely available to consumers when choosing coverage, and when they are, it is often in the form of technical documents that require specialized knowledge to parse. Therefore, to assist consumers, CMS created the MA Star Ratings.

Displayed next to the enrollment button in Medicare's unified shopping platform, the Star Ratings provide a coarse summary of the quality of each plan.[13] To compute the ratings, CMS first collects information on over sixty measures of quality for each plan and categorizes them into five groups: Outcome (e.g., readmission rate), Intermediate Outcomes (e.g., diabetes management), Access to Care (e.g., management of appeals), Patient Experience (e.g., customer service), and Process (e.g., breast cancer screenings). Having collected the data, CMS assigns a discrete measure-level score of one to five to each plan-measure, ascending in quality. Next, CMS computes a continuous score for each plan by choosing a weight for each category and computing a weighted average of all measure-level scores for each plan. Overall, denoting $\mathcal{K}$ the set of quality categories, and w_k the weight that each category $k \in \mathcal{K}$, and $\mathcal{L}_k$ the measurements included in the category, the rating of plan j is given by

$$\text{Rating}_j = \text{Round}_{.5}(\underbrace{\frac{\sum_{k\in\mathcal{K}} w_k \sum_{l\in\mathcal{L}_k} \text{MeasureScore}_l(q_{lj})}{\sum_{k\in\mathcal{K}} w_k|\mathcal{L}_k|} + \omega_j}_{\text{Continuous Score}}) \tag{1}$$

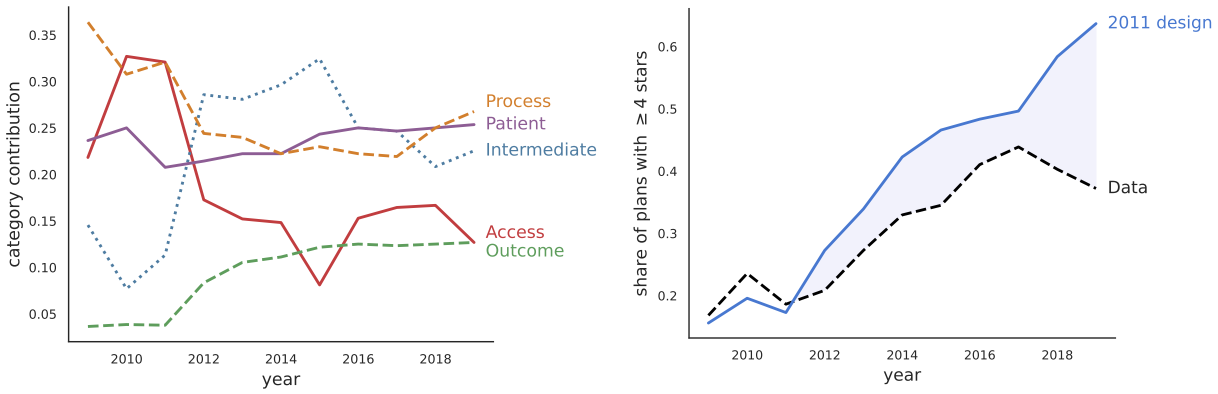
Where $\text{Round}_{.5}(\cdot)$ rounds a number to its nearest half and q_{kj} is the quality of plan j in measure k. In the expression ω_j is a combination of several regulatory features which I call the adjustment factor. I provide further details about the exact design in Appendix 3.1.1.[14]

CMS has frequently changed the weights and number of measures in each category, introducing substantial variation in the Star Rating design.[15] In 2012 CMS moved from uniform weights to a design that gives each Outcome and Intermediate Outcome measure three times the weight of any Process measure and twice the weight of any Access or Patient Experience measure. The size of each category changed each year as CMS introduced and removed measures from each. Overall, each category's contribution to the rating has varied significantly, as shown in Figure 2.a. As I detail Appendix 3 , this design variation was likely observed by consumers, as the composition of

[13] See Appendix figure 1 for a view of the platform. The Star Rating program has evolved in its form over the years. For a description of earlier designs, see Dafny and Dranove (2008).

[14] CMS computes the ratings at the contract instead of the plan level. Contracts are aggregations of multiple plans of the same insurer that share the same quality. Approximately, contracts define the network and quality of an insurance product while a plan determines the cost-sharing attributes of the product. All the terms entering equation (1) are at the contract level.

[15] CMS also varied the measure-level scoring function over the years, although their yearly change is modest. Appendix 3.1.1 provides further details.



(a) Design variation

(b) Variation in rating assignment

Figure 2: Scoring design variation and simulated effect under fixed quality

Notes: Figure (a) shows the change in the contribution of each category to the overall score. The contribution of a category corresponds to the product of its number of measurements (e.g., Process includes breast cancer screening and kidney disease monitoring) and its weight, divided by the total weight among all measurements. Figure (b) shows the change in rating assignment that would result if CMS had kept its 2011 scoring design, keeping quality as measured in the data. I select 2011 as the baseline year because that year's design shares a large number of measures with both previous and following years, making the comparison more natural. The shaded area highlights the gap across the resulting rating assignment. Adjustment factors are preserved as measured in the corresponding year.

categories was visible on the Medicare website and in booklets sent by CMS to enrollees.

This rating variation significantly impacted rating assignment. For example, increasing the share of enrollees receiving adequate care for their high blood pressure (an Intermediate Outcome measurement) from 45% to 65% was 86% more valuable for rating purposes in 2012 as in 2011. In contrast, increasing breast cancer screening (a Process measure) from 60% to 75% was 38% less valuable.[16] To illustrate the overall change in rating assignment, Figure 2.b shows that if CMS had kept the 2011 scoring design, 60% instead of 40% of 2019 plans would have received four or more stars. The gap between the two systems is largely due to a decrease in the importance of Access and an increase in that of both Outcome categories. Thus, in 2011 a high-quality plan was one that afforded consumers great access to physicians and a median-quality network of hospitals, while in 2019 the roles of hospital quality and access to physicians were reversed. The figures also show an improvement in overall quality as the share of top-rated plans increases regardless of the design.

Because insurers can offer the same network arrangement and services under different cost-sharing and premium combinations, CMS measures quality at a level slightly larger than a plan. A contract is a grouping of plans that share the same quality and insurer. The median contract has only two plans, with 70% of its enrollment in one of them. Consumers observe a rating for each

[16] Adequate care here means members with high blood pressure received treatment and were able to maintain a health pressure. Breast cancer screening rates are among women 40 to 69. The rating value is in terms of the continuous score.

plan and often will have only one of a contract's plan available in their county. Therefore, in many cases, the distinction between plan and contract is irrelevant. However, having price variation conditional on quality and year will prove useful for estimation purposes. Throughout, I refer to products in MA as plans and refer to contracts only when relevant for clarity or exposition.

Finally, CMS provides some dynamic incentives for quality. Starting in 2012, the rebate share and benchmarks of plans vary with their rating in the previous year, and the adjustment factor (ω_j) rewards quality improvements.[17] However, this paper aims to understand the short-run mechanisms, effects, and design of a purely informational quality disclosure policy. Thus, in the model and estimations that follow, I incorporate these dynamic features as they appear in the data and treat them as sources of revenue heterogeneity. I do not include pecuniary incentives as part of the designer's toolkit to avoid confusing gains from information design with those from direct transfers.

3.2 Data

This paper combines three data sources, the first being plan-market level data from 2009 to 2019. Each year, CMS publishes a compendium of data sets containing information on each MA plan in each county. I use it to construct a panel of plans with their market-level enrollment, benchmarks, prices, rebates, premiums, and the full detail of plan benefits and cost-sharing. Additionally, the data provides the total number of Medicare eligible beneficiaries in each county, and information regarding dual Medicare-Medicaid eligible population. I use these data to adjust the sample, removing dual eligibles and plans specifically designed for that population.[18] I present the descriptive statistics of these and the following data, together with further details regarding their construction in Appendix 3.2.

The second data source is the Medicare Current Beneficiary Survey (MCBS). This nationally representative rotating panel tracks around 15,000 Medicare beneficiaries each year, for up to four year. I obtain the data for 2009 to 2015, which provides me with information on individual demographic, well-being, income, location, and self-assessed knowledge about the Medicare program.[19] Most importantly, it provides plan choices that can be linked with the aggregate data. I restrict the data to non-dual beneficiaries, within the continental US, and with geographic information,

[17] MA also rewards plans achieving five stars by allowing consumers to switch into them after the open enrollment period ends. As there are few five-star plans, I exclude this behavior from the analysis by only considering demand within the open enrollment period. Another dynamic behavior not treated in this article is contract consolidation. This was a practice exploited by few insurers to combine contracts to manipulate the rating of their products for one year. I discuss this further in Appendix 3.1.

[18] Similar data restrictions have been used by Aizawa and Kim (2018), Miller et al. (2019) and Curto et al. (2021a).

[19] The MCBS for 2014 is not included as it was never released to the public because of implementation difficulties.

leaving 46,833 beneficiary-years. Importantly, the MCBS provides sampling weights to compare the survey's demographics with the national population. However, because of the limited size of these data, they do not include all counties. This will limit my final welfare measurement to about a third of the overall population of Medicare in 2015, or about 22 million individuals.

Finally, the third data source pertains to the quality of plans and the rating rules. Each year, CMS publishes the data used to compute the star ratings. These files contain CMS's quality measurements and their associated star rating and cutoffs. The challenge in using these data is that the measurement and scale of the underlying quality dimensions have changed over time. Additionally, the files do not contain the direction of improvement and range for dimensions measured but with zero rating weight ($w_k = 0$). To tackle this challenge, I complement the data by reviewing a decade of CMS communications to insurers regarding scoring design changes. From this, I fully recover the missing information on measures and uncover year-to-year changes to the scoring design. I review these rules in Appendix 3.1.1.

4 Demand and Quality Responses to Scoring

The score's effect in the monopoly regulation example of Section 2 relied on two fundamental market behaviors. First, consumers understood the rating implications and adjusted their beliefs accordingly. Second, the monopolist responded to the incentives created by the change in rating, and adjusted its quality. This section explores whether the rating changes in MA produced similar responses among consumers and insurers.

4.1 Demand Responses

The regulatory power of quality disclosure stems from its effect on demand. Consumers' responses can be decomposed into two margins: changes in levels, as a plan's rating changes over time; and changes in compositions, as the scoring design changes while keeping the awarded rating fixed. Responses on both margins are a natural consequence of consumers' choices assuming that they prefer higher quality (level) and care more about some categories of quality than others (composition).

The response of MA consumers to level changes in the score is well documented (Dranove and Dafny, 2008; Reid et al., 2013; Darden and McCarthy, 2015). To contribute to this evidence, I leverage my individual-level data and control for two potentially confounding effects. First, plans might be heterogeneous in dimensions beyond their financial characteristics and rated quality. If a plan's quality is positively correlated with consumers' unobserved preference for it, a simple analysis might inflate the effect of ratings on demand. I control for this source of bias by studying demand changes within a contract. Second, consumers in MA have substantial switching costs (Nosal, 2011), potentially due to the hassle of changing primary care physicians and interruption in treatments. Thus, some consumers might fail to switch away from a plan whose quality is dropping, which would bias the effect of ratings downwards in an analysis that uses aggregate market shares to measure demand. I control for this confounding factor by restricting the analysis to consumers not previously in MA.

To study the level effects, I regress consumers' plan choices on plans' star ratings, product and consumers' characteristics, contract fixed-effects, and market fixed-effects, resulting in the following specification.

$$y_{ijt} = \sum_{r=1}^{5} \alpha_r \mathbb{1}\{r_{jt} = r\} + \gamma_{c(j)} + \mu_{m(i)} + \xi_t + \boldsymbol{x}_{ijt}\boldsymbol{\lambda} + \epsilon_{ijt}$$

Above, y_{ijt} indicates whether consumer i chose plan j in year t, $\mathbb{1}\{r_{jt} = r\}$ indicates if the plan's rating is r, and $(\gamma_{c(j)}, \mu_{m(i)}, \xi_j)$ are fixed-effects for the plan's contract, the consumer's market, and the year, respectively.[20] Additional controls, $\boldsymbol{x}_{ijt}$, include demographic variables, such as education, self-assessed health status, ethnicity, gender, disability status, and plan characteristics, such as Part D coverage and additional plan benefits, such as dental coverage.

The first column of Table 1 presents the estimates of α_r. The results show that improving a plan's rating significantly increases its demand. For example, an improvement from four to five stars roughly increases the choice probability of the plan by 0.6% relative to all options, which corresponds to 31.6% increase in demand relative to the same plan when it had four stars. The gap between these two numbers is caused by new MA enrollees having both a very large set of plans to chose from and a high probability of choosing TM. In this restricted sample, the average MA plan is chosen only 1.58% of the time. Nevertheless, the effect of rating levels on demand is significant and, most importantly, meaningful for the revenue of insurers.

Having documented consumers' responses to changes in rating levels, I now turn attention to compositional effects. To do so, I ask whether the level-effect varies over years in a way that is systematically correlated with the contribution of each category to the rating, shown previously in Figure 2.a. Specifically, I add the contribution of category k, W_{kt}, to the previous regression, where W_{kt} is defined as the total weight of all measurements category k, relative to the total weight of all

[20] As insurers can offer the same insurance contract under different cost-sharing combinations and labels, I do not control for the label-indicator of plan in γ_j, but rather for the higher-level contract name. Contracts are insurer specific and are the level at which quality is measured by CMS. To reduce the number of effects to estimate in this limited sample, I round half-stars up to their nearest integer.

measurements in all categories.[21]

$$y_{ijt} = \sum_{r=1}^{5} (\alpha_r + \beta_r W_{kt}) \mathbb{1}\{r_{jt} = r\} + \gamma_{c(j)} + \mu_{m(i)} + \xi_t + \boldsymbol{x}_{ijt}\boldsymbol{\lambda} + \epsilon_{ijt}$$

Columns II and III of Table 1 show that the estimated coefficients for β_r are statistically significant for all categories. The results indicate that a rating improvement was more valuable in later years of the sample when the system required larger increases in Outcome and Intermediate Outcome quality for an improvement. The results indicate that increasing the contribution of Outcome quality to the rating by 1% increases the choice probability gained by a plan improving from three to four stars by 0.1%. This improvement is roughly a 7% increase in the demand for the improving plan.

These compositional effects also serve to test whether consumers are informed of the scoring design. Under the hypothesis of consumers being completely ignorant of the design, the change in demand when a product increases in rating should be independent of how that rating was calculated. Finding that consumers value rating increases differently when certain categories were more represented in the design rejects this hypothesis. This finding also agrees with the institutional features of MA, as the weights are largely given by the number of measurements included in each category, which are visible to consumers. I present additional supporting evidence for the claim that consumers understand some coarse features of the design in Appendix 4.

4.2 Quality Responses

Quality responses to changes in the scoring design can have large welfare consequences. To explore the causal link between design and quality, I examine variation stemming from the introduction of new quality measurements to the score. For example, in 2018, CMS introduced the Process measure *Medication Reconciliation Post-discharge*. This measure corresponds to the share of members whose medication records the insurer updated within 30 days of a hospital discharge. Updating this list is a valuable task as it helps prevent medical errors due to changing medication and helps insurers keep track of high-risk drug utilization in their population. In 2017, when this dimension was measured but not incorporated into the score, the average contract reconciled medication lists about 20% of the time. In 2018, when the measure impacted the rating, the average shifted to nearly 60%.[22]

While there are only eight of these events in the data, they have several advantages relative to

[21] Formally, $W_{kt} = (|\mathcal{L}_{kt}| w_{kt}) / (\sum_{k' \in \mathcal{K}} |\mathcal{L}_{k't}| w_{k't})$.

[22] Appendix Figure 11 shows the distribution of quality in both years.

Table 1: Demand Responses to Scoring

	I		II		III	
Rounded star rating (α_r)						
2 stars	-0.000	(0.003)	-0.035***	(0.006)	-0.033***	(0.008)
3 stars	0.001	(0.001)	-0.036***	(0.004)	-0.027***	(0.003)
4 stars	0.007***	(0.002)	-0.043***	(0.004)	-0.030***	(0.003)
5 stars	0.012***	(0.003)	-0.036***	(0.006)	-0.023***	(0.005)
Rating category weight (β_r)						
2 stars			0.728***	(0.116)	0.205***	(0.047)
3 stars			0.770***	(0.077)	0.189***	(0.019)
4 stars			0.881***	(0.083)	0.219***	(0.021)
5 stars			0.831***	(0.094)	0.204***	(0.025)
Category			Outcome		Intermediate	
N	421606		421606		421606	
R^2	0.712		0.713		0.713	

Notes: This table displays the estimates of the demand response to rating and scoring design. Only two categories are shown for space, the full table is shown in Appendix Table 7. Observations are weighted by the MCBS sampling weights. The omitted category are new plans and plans that don't have star ratings due to insufficient enrollment in previous years. Standard errors in parentheses are heteroskedasticity robust. *p<0.05, **p<0.01, ***p<0.001.

the other sources of design variation.[23] First, CMS measured quality in these dimensions before and after their introduction to the rating. Second, CMS announced these changes to insurers without anticipation. Third, unlike changes in the contribution of categories to the score, the effects of these changes are likely to concentrate on the introduced measures.[24] Finally, these changes created different incentives for different plans. Those with high quality before the introduction stand to gain scores if their quality remains the same. In contrast, those with low pre-introduction quality will lose ratings, demand, and revenue, as the previous section shows. As the effect of quality on scores is bounded (i.e., each measure has a threshold after which there are no incentives to improve), and the impact of scores on demand is non-linear, high-quality and low-quality plans face different incentives.

These features suggest an analysis that compares the quality of introduced measures between plans of high and low pre-introduction quality. If firms do not respond to rating incentives, and

[23] For the full list of affected measures, see Appendix Table 6.

[24] Because quality dimensions are substitutes within a firm's score, simultaneous variation in weights can lead to inversely correlated changes between weights and qualities. An analysis that does not account for the shift in overall investment incentives might fail to find any systematic effect or even get the wrong sign.

whatever generated the heterogeneity in quality before introduction is not affected by the design change, then the gap in quality across groups should remain the same. If, instead, firms react to the change in design, then we would expect the gap between the groups to narrow. I capture this idea in the following triple-differences regression:

$$\underbrace{q_{ljt}}_{\substack{\text{normalized quality}\\ \text{measurement}}} = \underbrace{\sum_r \beta_r T_{lt} \mathbb{1}\{G_{lj} = r\}}_{\text{treated-post-treatment group}} + \underbrace{\gamma_{lj} + \mu_{lt} + \xi_{jt}}_{\text{pairwise fixed effects}} + \epsilon_{ljt} \tag{2}$$

where T_{lt} indicates if measure l was introduced at any year equal or preceding t, G_{lj} indicates groups of pre-introduction quality for plan j, and $(\gamma_{lj}, \mu_{lt}, \xi_{jt})$ are measure-plan, measure-year, and plan-year fixed-effects.

This regression can be interpreted as a treatment analysis. The treatment unit is a plan's quality measure, the treatment is the act of including it in the score (T_{lt}), and there are different treatment groups (G_{lj}). The regression controls for heterogeneous starting points for different units (γ_{lj}), and compares the evolution across measures and years (μ_{lt}). Finally, the analysis also accounts for trends in plans' overall quality (ξ_{jt}) caused, for example, by improvements in the network of providers for reasons other than scoring design. Thus overall, I examine the difference in quality changes across plans (j), within newly introduced measures (l), across years (t).

It is informative to think of the standard differences-in-difference analysis to understand the variation identifying β_r. In that case, the regression would include plan-measure fixed-effects (γ_{lj}), as it is the unit of treatment, and measure-year fixed-effects (μ_{lt}), as treatments are assigned over the years differently for different measures. In that case, we would include only T_{lt} and, assuming homogeneous treatments, the coefficient on it would capture the average change in quality across measures over time. The problem with this approach is that CMS did not randomize which measures to introduce. Quality in some of the treated measures was already trending upwards before the treatment. The third difference, comparing across plan groups, allows me to control this trend under the assumption that CMS did not change the system because it foresaw a relative change in quality across plans. I include the treatment groups and the year-plan fixed-effect to introduce this third difference, which accounts for changes in a plan's overall quality over time. Variation in quality differences across treatment groups over time within a measure identifies the coefficient of interest, β_r.

I implement the regression in the following way. First, I transform all quality measurements to a common scale by standardizing them using their mean and standard deviation across all years. Second, I drop plan-measures in the first and last quartiles of quality in the year before the introduction to avoid conflating the effects of bounded quality domains.[25] Third, I define treatment groups according to the predicted measure-level score of each plan-measure using the design of the year of introduction but the quality of the year before.[26] While measure-scores run from one to five, the censoring of pre-treatment quality drops groups one and five, and I use group four as the point of reference. Therefore, there are three levels of β_r, capturing changes in quality of plans that would have received a measure-score of r if they had not changed their quality, relative to the change in the quality of plans that would have received a measure-score of four. This approach uses CMS's design to assign treatment intensities. An alternative which I present in Appendix 4 uses quartiles of the pre-treatment quality distribution and finds similar results.

Table 2 shows the results of this analysis. The first column shows a differences-in-differences approximation, which does not include the plan-year fixed-effect. This version omits the heterogeneity across plan trends and overestimates the effect on those with the highest incentive to improve. The second column shows the triple-differences coefficients. As expected, the magnitudes decrease with the predicted rating as incentives to improve drop. After treatment, a contract that would have obtained a single star under its pre-treatment quality improves by 0.43 standard deviations more than a contract predicted to obtain four stars. This is about 40% of the pre-treatment quality gap between the first and fourth treatment group. Appendix 4 provides further details and supporting evidence for this analysis, showing the lack of pre-trends, event-study plots, and robustness to common concerns with staggered differences-in-differences and dynamic treatment effects (Goodman-Bacon, 2021; Baker et al., 2021).

Overall, the estimates of this exercise show three things. First, quality is variable and can change quickly. Second, not all plans are affected equally by changes in regulation. Third, plans that stand to lose more increase their quality more. These are valuable facts for designing a scoring system, as they inform the extent to which the planner can alter quality in the market through this informational lever. However, as the measures were not chosen randomly by CMS, the results do not speak to the effect of scoring a generic quality measure. It is possible that other measures are more challenging to adjust and less affected by scoring incentives. The variation needed to measure these effects exists in the data, as the rating rules change every year, yet additional structure is required to disentangle the joint evolution of quality incentives.

The following section presents the additional structure I use. I introduce a model of insurance and demand that rationalizes the scoring effects found in this and the previous section. The model allows me to leverage the design variation further to uncover the primitives that govern

[25] The domain of most quality measures is bounded. Therefore, low-quality plans can only improve, and high-quality plans can only worsen, and a failure to account for this would inflate the measurements of this analysis. By censoring quality, I likely err on the side of under-estimating the effect.

[26] Formally, $G_{lj} = \text{MeasureScore}_{lt}(q_{ljt-1})$ where t is the year in which measure l is treated.

Table 2: Quality Responses to Scoring

	Differences-in-Differences		**Triple-Differences**	
Predicted score (β_r)				
1	0.485***	(0.106)	0.428***	(0.108)
2	0.365***	(0.100)	0.316**	(0.104)
3	0.076	(0.053)	0.045	(0.051)
Contract measure FE	Yes		Yes	
Measure year FE	Yes		Yes	
Contract year FE	No		Yes	
N	167693		167693	
R^2	0.678		0.693	
Mean standardized quality	0.0319		0.0319	

Notes: The estimated effect is relative to plans predicted to obtain a measure-level score of 4. The dependent variable is standardized quality in each measure, relative to the mean and standard deviation across all years. To avoid boundary issues, the first and last quartile of pre-treatment quality are excluded. Standard errors, in parentheses, are clustered at the contract level. For further details see Appendix 4.

these effects. Specifically, I use variation in product characteristics across markets and time and a panel of consumer choices to recover preferences for insurance plan attributes. Using the scoring variation and the individual-level demand, I estimate consumers' willingness to pay for each score in each year. Changes in demand and subsidy rules perturb the marginal revenue of insurers, which I use to estimate their marginal cost of insurance. Finally, I exploit the evolution of investment incentives caused by the changing scoring designs to estimate insurers' cost of supplying quality.

5 Model

I model the MA demand and supply behavior as the Perfect Bayesian equilibrium of a game consisting of repeated static interactions between consumers and insurers. At the beginning of each year, the planner announces a national quality scoring rule. Insurers simultaneously invest in quality, which stochastically determines realized qualities. Insurers then choose their plans' prices, which subsidies and regulation convert to premiums and cost-sharing benefits. Finally, consumers observe premiums, benefits, and ratings, and choose whether to enroll in TM or in one of the MA plans available in their county. Figure 3 illustrates the game's timing and information, with bold letters denoting vectors.

The model simplifies several features of the market. First, insurers' quality investment problem is at the average category level, which reduces the computational and statistical burden associated

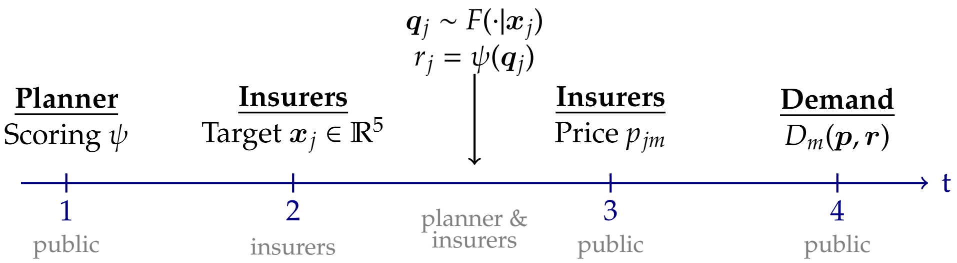

$\boldsymbol{q}_j \sim F(\cdot|\boldsymbol{x}_j)$
$r_j = \psi(\boldsymbol{q}_j)$
Planner
Scoring ψ
Insurers
Target $\boldsymbol{x}_j \in \mathbb{R}^5$
Insurers
Price p_{jm}
Demand
$D_m(\boldsymbol{p}, \boldsymbol{r})$
t
1
public
2
insurers
planner & insurers
3
public
4
public

Figure 3: Model Timing

Notes: j indexes plans and m markets. $\boldsymbol{x}_j$ denotes an insurers target quality, with $\boldsymbol{q}_j$ being its realization. r_j is the rating assigned to plan j given the scoring rule $\psi(\cdot)$. The gray text under each stage indicates who observes the choices in the given stage.

with measure-level choices. However, these are still multidimensional, which allows me to capture multitasking behavior. Second, as CMS has experimented with different designs, I impose no optimality conditions on their choices. I do not require the scoring variation to be random or exogenous to changes in aggregate preference for MA or insurers' cost of providing quality. Finally, to focus on information-based regulation, I hold fixed price and subsidy regulations and firm entry choices.

Next, I present the components for each stage of the model. I discuss the implications of the model's simplifications for the question of designing quality-regulating scores at the end of this section.

5.1 Demand

I model the demand for MA plans following Aizawa and Kim (2018) and Miller et al. (2019), but diverging in two aspects. First, I follow Curto et al. (2021a) and include the dollar value of cost-sharing benefits instead of the different deductibles and coinsurance rates that define it. Consumers observe a similar value in the platform and CMS regulates cost-sharing at this aggregate level, which makes the simplification appealing. Second, I model quality preferences explicitly which will become crucial when evaluating counterfactual scoring designs.

Each year t, consumers in county m are offered a collection of MA insurance plans $\mathcal{J}_{mt}$. Each plan is characterized by a total premium p^{total}_{jmt}, cost-sharing benefits level b_{jmt}, additional plan attributes $\boldsymbol{a}_{jmt}$ (e.g., bundled vision and dental insurance), and a star rating r_{jt}. Consumers maximize a Von Neumann-Morgenstern expected-utility, evaluating subjective beliefs over quality given observed product ratings and the scoring rule. The expected utility of consumer i of choosing plan j in market m at year t is given by:

$$u_{ijmt} = \underbrace{\alpha_i p^{\text{total}}_{jmt}}_{\text{premium}} + \underbrace{\beta_i b_{jmt}}_{\text{benefits}} + \underbrace{\mathcal{E}[v(\boldsymbol{q})|r_{jt}, \psi_t]}_{\text{quality}} + \underbrace{\boldsymbol{\lambda}^{a\prime}\boldsymbol{a}_{jmt}}_{\substack{\text{plan}\\\text{attributes}}} + \underbrace{\boldsymbol{\lambda}^{d\prime}\boldsymbol{dem}_{it}}_{\substack{\text{consumer}\\\text{demographics}}} + \underbrace{\boldsymbol{\lambda}^{l\prime}\boldsymbol{l}_{ijt}}_{\substack{\text{lock-in}\\\text{indicators}}} + \underbrace{\xi_{jmt}}_{\substack{\text{unobserved}\\\text{preference}}} + \underbrace{\varepsilon_{ijmt}}_{\sim T1EV} \tag{3}$$

Consumers have heterogeneous preferences for premiums and benefits (α_i, β_i), and value quality according to the subjective expectation of a common function ($\mathcal{E}[v(\boldsymbol{q})]$) given the rating of the plan (r_{jt}) and the current scoring system (ψ_t). For the time being, I impose no further assumptions on how consumers arrive at this posterior belief or the function $v(\cdot)$. Instead, I leverage that CMS groups plans of the same quality under the label of a contract, which allows the quality-utility term to be captured as a contract-year fixed-effect. Plan choice is also affected by other bundled services ($\boldsymbol{\lambda}^a$), preference heterogeneity for MA versus TM across demographic groups ($\boldsymbol{\lambda}^d$), and by previous relationships with firms and products in the market ($\boldsymbol{\lambda}^l$). Following Handel (2013), this last factor captures MA enrollment inertia (Nosal, 2011) as a direct utility impact. Finally, consumers have time-varying unobserved preferences for plan-markets (ξ_{jmt}) and an independent random type-1 extreme preference shock perturbs the utility of each choice (ε_{ijmt}).[27]

Consumers also have the option of choosing TM coverage. Given that the vast majority of MA consumers choose plans with bundled prescription drug coverage, I assume that the relevant outside option for the population corresponds to a bundle of TM and stand-alone part D coverage. I denote b_0 as TM's standard insurance benefits, and p^D_{0mt} as the price of the most popular part D plan in county m and year t. The utility of the outside option is given by

$$u_{i0mt} = \alpha_i p^D_{0mt} + \beta_i b_0 + \varepsilon_{i0mt} \tag{4}$$

Given this model, the expected demand for product j in market m in year t is the sum of the probabilities with which each consumer chooses the product.

$$D_{jmt} = \sum_{i \in \mathcal{I}_m} s_{ijmt} = \sum_{i \in \mathcal{I}_m} \underbrace{\frac{\exp(u_{ijmt} - \varepsilon_{ijmt})}{\exp(u_{i0mt} - \varepsilon_{i0mt}) + \sum_{j' \in \mathcal{J}_{mt}} \exp(u_{ij'mt} - \varepsilon_{ij'mt})}}_{\text{individual choice probability}}$$

[27] Consumers in MA and TM must pay a fixed part B premium. As this premiums is common across all options, I normalize it in this exposition.

5.2 Supply

5.2.1 Insurers' pricing problem: Each year t, at the third stage of the game, insurance firm f observes the vector of realized qualities $\boldsymbol{q}_t$, and its associated ratings vector $\boldsymbol{r}_t$. Given this information, the firm chooses prices to maximize its total profits.[28]

$$V_{fmt}(\boldsymbol{q}_t, \psi) = \max_{\{p_{jmt}\}_{j \in \mathcal{J}_{fmt}}} \sum_{j \in \mathcal{J}_{fmt}} \underbrace{D_{jmt}(\boldsymbol{p}_{mt}, \boldsymbol{r}_t)}_{\text{demand}} \gamma_{jmt} \left(\underbrace{p_{jmt} + R(p_{jmt}, \boldsymbol{z}_{jt})}_{\text{marginal revenue}} - \underbrace{C(\boldsymbol{q}_{jt}, \boldsymbol{a}_{jmt}, \boldsymbol{\theta}^c)}_{\text{marginal cost}} \right) \tag{5}$$

In this equation, $D_{jmt}(\cdot)$ is the plan's aggregate demand as a function of the vectors of prices and ratings. This demand is multiplied by the risk-adjustment factor γ_{jmt}, which CMS determines for the plan before pricing or demand are realized. The marginal revenue of the plan is the sum of its price and additional revenue sources $R(\cdot)$. This second source is, in part, the result of the market's regulation and depends on the plan's price (p_{jmt}) and other plan attributes ($\boldsymbol{z}_{jt}$) which include its counties of service, prescription drug coverage prices, and the way in which the firm allocates certain subsidies into consumer benefits, among others. I present the full formula for this function and the way in which prices map to premiums and benefits in appendix 5. The cost of covering each enrollee's standard Medicare benefits, prescription drugs, and any non-Medicare extra benefits (e.g., dental insurance), as well as management costs, are contained in $C(\cdot)$. This function varies according to the plan's quality and additional attributes, and a set of unknown parameters to estimate $\boldsymbol{\theta}^c$, which are the only unknowns of this stage of the model.

The market's price and benefit regulation introduce a kink in the demand and revenue of a firm as a function of prices. If the firm sets prices above the kink, called the plan's benchmark, then a dollar increase in prices translates to an equivalent increase in revenue and premiums, and cost-sharing is not affected. Below the kink, a dollar increase in prices translates into less than a dollar increase in revenue and premiums and a mandatory decrease in the cost-sharing benefits of the plan, in an amount not exceeding a dollar.

5.2.2 Insurers' investment problem: In the second stage of the game, each firm observes the regulator's chosen scoring rule ψ_t and chooses an investment level x_{ckt} for each of its contracts c

[28] The price of a plan in MA is often called its bid. I avoid this terminology to prevent confusing this market's organization with an auction.

and category of quality k to maximize its total expected profits.[29]

$$\pi_{ft}(\psi_t) = \max_{\boldsymbol{x}_{ft}} \sum_m \underbrace{\int \mathbb{E}_{mt}[V_{fmt}(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)]dF(\boldsymbol{q}_f|\boldsymbol{x}_{ft})}_{\text{expected insurance profit}} - \underbrace{I(\boldsymbol{x}_{ft}, \boldsymbol{\mu}_{ft})}_{\text{investment cost}} \tag{6}$$

The firm's total expected profits are equal to its expected insurance profit ($V_{fmt}(\cdot)$) in each market m minus the cost of the quality investment ($I(\cdot)$). To derive an expectation for its insurance profits, the firm evaluates two dimensions of uncertainty. First, realized quality might differ from its intended target, which the model captures using a conditional distribution $F(\boldsymbol{q}_f|\boldsymbol{x}_{ft})$ where $\boldsymbol{x}_{ft}$ is the full vector of the firm's investments across all contracts and categories. Second, firms have some uncertainty about the actions of their rivals ($\boldsymbol{q}_{-f}$), which is represented by the expectation operator $\mathbb{E}_{mt}$. The investment cost is a known function of the firm's choices and some unknown parameters $\boldsymbol{\mu}_{ft}$. These cost parameters together with the conditional distribution that maps investments to quality are the two unknowns of this stage of the model.

Realized quality and investment targets are connected by the relationship $q_{ckt} = \Phi_k(x_{ckt} + \epsilon^M_{m(c)kt} + \epsilon^F_{f(c)kt})$, with $\Phi_k(\cdot)$ being a known strictly increasing function. There are two independent (from each other and the target) mean-zero errors in this expression. The first, $\epsilon^M_{m(c)kt}$, captures market-level shocks to a contract's quality and is common across all other contracts offered in the same market. This term captures population shocks that distort quality supply, such as a harsh flu season or a community vaccination drive. The second, $\epsilon^F_{f(c)kt}$, captures unexpected firm-level deviations in the production of quality, such as cost shocks when negotiating contracts with providers, firm-level congestion in following up with patients, or common failures in contracts written to promote quality at the physician level. The sum of these shocks represents that insurance firms have imperfect control over the quality of their plans. Instead, they form networks and write contracts that attempt to achieve certain targets but might fall short of or exceed the firm's intended goals. I provide empirical evidence indicating that MA insurers do not perfectly regulate quality in Appendix 3.1. Given this structure, the distribution of quality conditional on investment ($F(\boldsymbol{q}|\boldsymbol{x})$) is fully specified as a function of investments, the distribution of quality shocks (ϵ^F, ϵ^M), and the set of mappings Φ_k.

Insurance firms are secretive about their contractual arrangement with providers. Quality investments and targets are not observable in any common data source, even ex-post. Because of this, I follow the strategy of Sweeting (2009) and assume that firms hold rational expectations over the distribution of rival targets, formed through observation of market characteristics at investment time. Particularly, I assume that firms know the identity of their rivals in each market,

[29] The set of firm contracts C_{ft} and plans $\mathcal{J}_{ft}$ are related in the following way. Each contract is associated with a set of plans $\mathcal{J}_{ct}$ and $\mathcal{J}_{ft} = \bigcup_{c \in C_{ft}} \mathcal{J}_{ct}$.

the demographic characteristics of consumers, and their previous contract choices.

5.3 Discussion

The model makes two key simplifications that might affect the scoring design analysis. First, consumers have homogeneous preferences for quality. I make this simplification for computation reasons. While the model and methodology that follows could accommodate observable heterogeneity in quality preferences – maintaining identification and overall procedures – the computational cost of solving the scoring design problem with heterogeneity is gargantuan. The method I develop addresses a functional optimization problem over a non-smooth functional space and depends on a pre-computation that suffers from a curse of dimensionality. Adding a quality preference group would increase the time required to solve this problem from months to years.

Nevertheless, the loss from this simplification is unlikely to be meaningful for the MA case. Given that quality is a vertical attribute, any heterogeneity would consist of some population groups preferring some quality dimensions over others. The alternative design I develop increases quality across the board, making all consumers better off, regardless of their preferences. Moreover, I will show that quality-regulating scores can be designed even without knowledge of consumer preferences, at some loss of optimality.

The second key simplification is that the game is static. Consumers do not learn from their past experiences with insurers, and firms do not carry on investments from their previous years. However, quality in MA varies rapidly, as it is primarily the outcome of contractual arrangements. The variation I document in the settings and descriptive evidence sections supports this claim. Moreover, the largest insurers in MA have been in existence for decades and probably already invested in major components such as developing relationships with providers or software to track their populations' health. Therefore, dynamic investment incentives are likely of second order. For consumers, the problem is statistical, as significant inertia hampers the separate identification of learning from switching cost. The rotating-panel structure of the MCBS further complicates this, as it follows consumers for only a few years. Regardless, the data does not suggest significant differences among MA insurers in their ability to produce quality. Compounded with significant quality variations, it is improbable that information acquired in a given year will be valuable the next one. The model does, however, allow consumers to have systematic prferences for certain insurers or contracts, which might capture some long-run learning about their baseline unrated quality. Finally, only consumers with severe health complications are likely to learn the more nuanced quality dimensions (e.g., hospital network quality). They are also likely to be the least affected by a change in scoring design due to the switching costs associated with ongoing treatment or illness.

6 Identification and Estimation

This section discusses how the data identifies the model presented above and how I estimate its unknown components. I formalize two identification arguments: the non-parametric identification of the distribution that maps quality investments into realizations; and the semi-parametric identification of consumers' quality preferences and beliefs. For the sake of exposition, I remove the formal statements and proofs of these arguments to Appendix 6. This appendix also presents the technical details about the implementation and estimation steps.

6.1 Demand

I estimate the demand model using the two-step estimator of Goolsbee and Petrin (2004). The first step, which combines a weighted maximum likelihood criterion with a nested fixed-point routine, recovers preference heterogeneity and aggregate mean valuation for each product-market-year. The second step, consisting of a two-stage least-squares, decomposes the aggregate mean valuation into its components while accounting for price and benefit endogeneity. Together, the estimation steps combine the information contained in the limited individual-level choice data with market-level market shares for the universe of products. Individual choices are used to recover heterogeneity across consumers in preferences for product attributes, while product-market level data delivers common preferences.

The first step of the estimation separates consumers' utility (equation (3)) into individual-specific and common preferences. This is done by first splitting the premium and benefit preferences (α_i, β_i) into their mean (α, β) and variation $(\tilde{\alpha}_i, \tilde{\beta}_i)$. Next, all common components of the utility of product j at market m in year t are added into a single scalar δ_{jmt}. This component sums mean preferences for premiums, benefits, plan attributes, and quality. Thus, the first stage has five unknown components to estimate: preference heterogeneity in premiums and benefits $(\tilde{\alpha}_i, \tilde{\beta}_i)$, the effect of consumer demographics $(\boldsymbol{\lambda}^d)$, the effect of switching costs $(\boldsymbol{\lambda}^l)$, and a series of plan-market-year fixed effects $(\boldsymbol{\delta})$. Letting $\boldsymbol{\vartheta}$ denote the collection of these unknown components, the first-stage estimator solves

$$\max_{\boldsymbol{\vartheta}} \quad \underbrace{\sum_t \sum_i w_{it} \sum_{j \in \mathcal{J}_{m(i)t}} y_{ijmt} \ln(s_{ijmt}(\boldsymbol{\vartheta}))}_{\text{weighted log-likelihood}} \qquad \text{s.t} \quad \underbrace{s^*_{jmt} = \sum_i w_{it} s_{ijmt}(\boldsymbol{\vartheta})}_{\text{share matching}} \quad \forall j, m, t \tag{7}$$

Where y_{ijmt} indicates that consumer i chose plan j in the respective county-year, $s_{ijmt}(\boldsymbol{\vartheta})$ is the model-implied individual choice probability for the same, and s^*_{jmt} is the observed market share of the associated product-market. I weight the likelihood and constraints by w_{it} to adjust the MCBS sampling frequencies to represent the national population. I solve this maximization problem by optimizing over all elements except for the plan-market-year fixed effects, recovering the latter using the Berry (1994) inversion and the Berry et al. (1995) fixed-point equation.

The second step of the estimator is a two-stage least-squares (TSLS) regression of the estimated mean preference for products ($\hat{\boldsymbol{\delta}}$) into its components, recovering all remaining utility parameters. Firms' knowledge of consumers' unobserved preferences for products (ξ_{jmt}) when pricing creates a correlation among the second-stage residual, premiums, and benefits. To address this endogeneity, I develop instruments based on regulatory features of insurers' additional revenue ($R(\cdot)$). First, I leverage variation in the kink of R across plans and years. The kink's location depends on TM's cost in every county in which the plan participates, suggesting using TM's cost in every other market in which the plan operates as an instrument.[30] By construction, this instrument is unlikely to express any systematic preference for a specific MA plan as it is associated with the outside option's cost in other markets. Moreover, the second stage includes market and contract-year fixed effects, limiting the residual unobserved preferences.[31] The second instrument uses variation across plans in the added revenue they obtain when pricing below the kink.[32] While only the plan's price is endogenous, this second variable helps distinguish between its effect on premiums and benefits. As the second stage includes year-contract fixed effects to capture quality preferences, this instrument varies only across plans due to county choices. Both instruments have yearly variation caused by changes in regulation and TM's cost. Appendix Table 16 presents the first-stage estimates for these instruments.

6.1.1 *Quality beliefs and preferences:*

The previous estimation step recovered consumers' valuation for ratings ($\mathcal{E}[v(\boldsymbol{q})|r_{jt}, \psi_t]$) as contract-year fixed effects. The entire model can be estimated without imposing further structure on this component. However, counterfactual changes in the scoring design (ψ_t) will affect this object, as consumers will see new ratings assigned under new rules. This demands additional structure on rating valuations, sufficient to compute them under counterfactual scenarios.

The standard assumption in the literature is that consumers understand the scoring design and derive posterior expectations through the Bayesian updating of some prior. For example, the common assumption in the Bayesian persuasion literature is that the signal receivers (consumers) have accurate priors over the state and understand the signal structure (Kamenica and Gentzkow,

[30] Specifically, the first instrument consists of the leave-one-out average of market-level benchmarks for each plan-market-year, excluding the current market.

[31] Failure of the exclusion restriction would require, for example, plans to change counties as the correlation between TM cost and plan-specific preference varies. As 92% of non-terminated plans remain in a county the following year, this concern seems unlikely.

[32] This instrument corresponds to the rebate fraction for plans pricing above the benchmark and one for the rest.

2011). In the empirical literature, several papers have used parametric models of Bayesian demand assuming that signal receivers understand its structure (Crawford and Shum, 2005; Dranove and Sfekas, 2008; Chernew et al., 2008; Brown, 2018; Jin and Vasserman, 2019; Barahona et al., 2020).[33] I rely on a similar assumption for the main counterfactual analysis. As supported by the descriptive evidence, I assume that consumers understand how the scoring system partitions the space of qualities at the category level. That is, consumers know that in year t a plan receiving a rating r must have its average quality in each category within a particular set $\boldsymbol{Q}_{rt}$ (see Appendix Figure 3 for an illustration). The fact that ratings in MA are well represented as partitions of the space of qualities is a helpful feature that stems from CMS's choices of categories and weighting schemes.[34] It is also an inherent feature of partitional scoring designs, such as quality certifications.

This assumption, which I call *informed choice*, requires consumers to know the contribution of categories to the rating and certain cutoff values that determine where one rating ends, and a new one begins. Variation in category contribution is largely due to changes in the number of measures composing each category, which consumers can observe in CMS's platform. The cutoffs depend on the measure-level scores, which change only moderately from year to year, allowing some learning to occur. In total, the assumption states that consumers' hold some prior over quality f, which they update after observing a rating r, to compute the expectation of $v(\boldsymbol{q})$ conditional on $\boldsymbol{q}$ being within a set $\boldsymbol{Q}_r$. The assumption does not require the prior to be accurate and does not impose any parametric structure on the distribution of signals, relying instead on its true features.

One might wonder if this assumption, together with data on plan choices, is enough to identify consumers' quality preferences ($v(\cdot)$) separately from their beliefs about it (f). Naturally, consumers' valuation for ratings is identified from their willingness to trade premium increases for scores. However, is it possible to tell if consumers are willing to pay more for plans with five stars than those with four because they believe the quality difference is small yet valuable or large but less valuable? It turns out that the answer is yes. Assuming that MA ratings would continue to vary as they have done so far and that $v(\cdot)$ is linear, I show in Appendix Thereom 1 that choice data separately identifies preferences from beliefs, without imposing any parametric restriction on the prior .[35]

The intuition behind this result is that consumers' willingness to pay for rating increments implies bounds on their preferences and beliefs. For example, suppose quality is scalar, the prior is

[33] Alternatively, they assume that the consumers interpret the signal as if it originates from a specific and known parametric signal linked through moments to the data.

[34] Appendix 6.1 shows how MA ratings can be reconstructed at this level with minimal loss.

[35] An alternative, fully non-parametric argument, would note that ratings define lotteries over qualities and appeal to the result of Anscombe and Aumann (1963). However, this argument would require ratings to cover the entire space of lotteries, which is unrealistic for the limited type of rating variation in MA. The result I derive only uses variation consistent with the data.

uniform, and $v(q) = \gamma q$ with $\gamma = 1$. If there are nine scores uniformly dividing $[0, 1]$, then consumers would be willing to pay $8/9$ more for a top-rated product ($q \in [8/9, 1]$) than for bottom-rated on ($q \in [0, 1/9)$). Some simple algebra shows that by observing the differences in willingness to pay, and knowing the scoring structure, γ can be bounded within $(8/9, 8/7)$. Scoring variation produces new intervals for γ, which intersect and shrink the identified set down to a point. This structure also bounds the posterior beliefs consumers can hold, and thus their priors. This example and its formal counterpart rely on the identification of consumers' valuation for rating-years. Berry and Haile (2020) provide relatively general conditions for identifying such systematic preferences from individual-level choice data. Importantly, because of this result, my proof does not rely on the logit structure.

The result shows that informed choice is a powerful assumption. It imposes a strong structure on consumers' understanding and, in return, delivers identification. However, it is not strictly necessary in order to design a quality-regulating score. In Section 8, I replace the informed choice assumption with one that assumes consumers are entirely uninformed of changes to the scoring design. In the appendix, I show that this assumption implies that preferences are only set-identified, proving that lacking assumptions on how consumers interpret ratings, choice data alone does not identify preferences and beliefs over quality. However, the regulator can still bound the worst-case scenario for consumers' preferences and design a system that improves quality.

Overall, the results of this section deliver two key observations for scoring design. First, if consumers do not understand the scoring system, it is impossible (without ad-hoc assumptions) to identify their preferences for quality. This failure also impedes the regulator from assessing the optimality of its design. This observation might explain why successful disclosure systems are often accompanied by a public information campaign, as consumers' understanding of the system is helpful to them and the regulator. Second, consumers do not need to understand the scoring design for improvements to exist or be attainable. This statement, of course, depends on how ineffective the status-quo design is.

For the results that rely on the informed choice assumption, I estimate preferences (now captured by a vector $\boldsymbol{\gamma}$), and prior beliefs ($f(\cdot)$) using a non-parametric minimum distance estimator. To remove any systematic preference for specific contracts, I only leverage time-series variation within the contract's valuation, $\eta_{c(j)t} \equiv \mathcal{E}[\boldsymbol{q}|r_{c(j)t}, \psi_t]$, which I recover in the second stage of the main demand estimator. The resulting estimator is

$$\min_{\gamma,\zeta} \sum_{c(j)} \sum_{t} \sum_{\tau > t} \left(\Delta_t^\tau (\eta_{c(j)t} - \boldsymbol{\gamma}' \mathcal{E}[\boldsymbol{q}|r_{c(j)t}, \psi_t; \zeta]) \right)^2 \tag{8}$$

where $\Delta_t^\tau x_t \equiv x_\tau - x_t$ is the time difference operator and $\boldsymbol{\zeta}$ corresponds to the Fourier-coefficients of a series expansion of the common prior $\boldsymbol{f}(\cdot)$ onto a Fourier series. This step does not affect other estimates and can be safely disregarded when relying on the assumption of ignorance.

6.1.2 Estimates: Table 3 presents the main demand estimates. Panel A shows the estimated preferences for premium and benefit levels. A dollar in benefits is roughly equivalent to a two-and-a-half dollar reduction in premiums for a low-income male of "fair" perceived health. Poorer and healthier consumers are more responsive to premiums and benefits. The distaste for premiums decreases with age but benefits preferences are inverse-u-shaped, peaking between 70 and 75. The average price elasticity – a statistic that aggregates premium and benefits preferences – is -8.34. As a comparison, a single-product monopolist with constant marginal cost and no part D coverage would set prices to meet an elasticity of -1, assuming a low enough benchmark.[36] Panel B presents some of the additional preferences for fixed product attributes. Importantly, consumers have a strong preference for products bundling prescription drug coverage. Appendix Section 6.2 presents the full set of estimated coefficients, including the large switching costs across systems, insurers, and contracts.

Panel C of Table 3 presents consumers' quality preferences under the assumption of informed choice. Consumers value all dimensions of quality but at different magnitudes. The most valued category is Outcomes, with maximal quality begin valued at $4498 in yearly premiums. In contrast, Intermediate Outcomes is the least valued category, with maximal quality valued at $1654. However, quality dispersion differs across categories. A standard deviation improvement in Outcome is worth $204, as quality dispersion is low. In Intermediate, a standard deviation improvement equals $194, as the dispersion is large.

Using these estimates and following Train (2015), I compute the surplus loss from consumers' incomplete information about product quality holding product attributes fixed. I estimate that the average Medicare beneficiary loses approximately $185.9 per year relative to full information. This amount – equal to nearly six months of average MA (part C) premiums – indicates the potential gains from increasing the informativeness of the score. Appendix Section 6.2 presents additional details about this exercise and the estimated consumer beliefs.

6.2 Supply

I estimate the different components governing price and quality competition in three steps. First, I use optimality conditions from the pricing problem to estimate the marginal cost of insurance. Second, I use a high-dimensional non-parametric conditional density estimator to recover the quality

[36] The table also provides premium elasticities as a comparison with the previous literature. For example, Miller et al. (2019) estimate a premium elasticity of -2.6 using similar data but a different model. Such a high premium elasticity would imply excessive price elasticities leading to negligible firm markups in my model.

Table 3: Key Demand Estimates

Panel A: Premium and benefit mean preference and heterogeneity

	Premium (α_i)		**Benefits (β_i)**	
Mean preference	-1.112**	(0.393)	2.915***	(0.383)
Medium income	0.041	(0.057)	-0.028	(0.071)
High income	0.271***	(0.060)	-0.167*	(0.073)
Female	-0.057	(0.046)	-0.006	(0.058)
Age group < 65	-0.111	(0.091)	-0.005	(0.099)
Age group $\in [70, 75)$	-0.009	(0.057)	0.151***	(0.041)
Age group $\in [75, 85)$	0.017	(0.055)	0.102*	(0.040)
Age group ≥ 85	0.195*	(0.081)	-0.110	(0.058)
Health - Excellent	-0.262***	(0.077)	0.010	(0.057)
Health - Very Good	-0.226**	(0.069)	-0.022	(0.052)
Health - Good	-0.132	(0.068)	-0.044	(0.050)
Health - Poor	-0.005	(0.116)	-0.145	(0.083)

Panel B: Other product attributes (λ^a)			**Panel C: Quality preferences (γ)**		
Drug deductible	-0.001***	(0.000)	Access	4.501***	(0.365)
Part D coverage	1.778***	(0.020)	Intermediate	1.839***	(0.042)
Dental cleaning	1.846***	(0.060)	Outcome	5.002***	(0.807)
Hearing aids	-0.229***	(0.031)	Patient	3.792***	(1.112)
Vision insurance	-0.032	(0.023)	Process	2.315***	(0.161)

Panel D: General information

Observations	36447	Weighted log. likelihood	-5.131
Mean price elasticity	-8.348	Mean premium elasticity ($p^C > 0$)	-0.951

Notes: Panel A and B report key estimates of individual preference for product attributes, corresponding to equation (3). In Panel A, the omitted category is low income males of "fair" self-reported health-status. Income groups are defined by terciles of the MCBS income distribution. Premiums and benefits are measured in thousands of dollars per year. Panel C reports quality preference estimates under the assumption of informed choice. All observations are weighted by the MCBS sample weights. Un-adjusted heteroskedastic standard errors in parenthesis. *p<0.05, **p<0.01, ***p<0.001.

shock distribution. Finally, I compute the expected marginal profits from quality investment to estimate the investment cost. I discuss each in order.

6.2.1 *Insurance marginal costs:*

The first-order optimality conditions (FOC) of the insurer's pricing problem equates marginal revenue with marginal costs.[37] As the marginal revenue of a firm depends on its observed demand, prices, and estimated demand elasticity, this provides an opportunity to recover the firm's marginal cost parameters ($\boldsymbol{\theta}^c$). I assume that the marginal cost

[37] As the firm's problem is not differentiable at the benchmark, the FOC is only valid for prices away from this cutoff. However, as in the data no firm violates this conditions, this condition must holds for all observed prices.

function is linear, leading to the following expression for any firm f:

$$\underbrace{\boldsymbol{p}_f + R(\boldsymbol{p}_f, \boldsymbol{z}_f)}_{\text{revenue per consumer}} \; + \underbrace{(\nabla \tilde{\boldsymbol{D}}_f')^{-1}(I + \nabla R_f(\boldsymbol{p}_f, \boldsymbol{z}_f))\tilde{\boldsymbol{D}}_f}_{-\text{profit margin}} = \underbrace{\boldsymbol{\theta}_q^{c\prime}\boldsymbol{q}_f + \boldsymbol{\theta}_a^{c\prime}\boldsymbol{a}_f + \boldsymbol{c}_f}_{\text{marginal cost } = \boldsymbol{C}(\boldsymbol{q}_f, \boldsymbol{a}_f, \boldsymbol{\theta}^c)} \tag{9}$$

where gradients are all with respect to the vector of prices $\boldsymbol{p}_f$, and $\tilde{\boldsymbol{D}}_f$ is the risk-adjusted demand vector. The identity states that revenue per consumer minus the firm's profit margin equates the marginal cost. The margin depends on consumers price elasticity and the change in additional revenue produced by CMS's price regulation. On the right-hand side, I have decomposed the firm's marginal cost into its quality components ($\boldsymbol{q}_f$), its systematic observable components ($\boldsymbol{a}_f$), and its residual plan-market-year specific component ($\boldsymbol{c}_f$).

Variation in demand, competition, and regulation all serve to identify marginal costs. Substituting in the demand estimates in the optimality condition above delivers the marginal cost estimates. Panel A of Table 4 presents the estimates of $\boldsymbol{\theta}_q^c$, when $\boldsymbol{a}_f$ includes contract, year, and market fixed effects, as well as controls for all additional bundled services provided by the plan. Therefore, the effect of quality on marginal costs is determined by assessing how a plan's marginal revenue varies when its quality changes in ways that are not common to the market or the national trend in quality. The estimates indicate that improving both types of medical outcomes increases marginal cost. In contrast, improving Process and Patient quality lowers marginal cost. One justification for this reversal is that Process includes preventive care and the management of expensive chronic illnesses. For an elderly population, these can help prevent expensive hospitalization and reduce costs (Newhouse and McGuire, 2014). Having better physicians in the network – captured by Patient quality – is likely associated with similar improvements and might make patients more likely to adhere to preventive and diagnostic care. Of course, these negative effects do not imply that firms should set these qualities to their maximum, as there might still be significant investment costs.

These estimates imply reasonable markups for insurers. The average plan markup is 11.2%, while for the top 4 insurers, it is 13.3%. As a point of comparison, Curto et al. (2019) use the Health Care Cost Institute data to estimate that in 2010, the average insurer in their sample spent \$590 per enrollee risk-month in medical costs, or \$680 in adjusted 2015 dollars. My estimate for the same set of firms is an average of \$771, including medical and administrative costs. This comparison suggests that about 10% of marginal cost is administrative, which is coherent with the level of involvement of MA insurers with their enrollee's health.[38]

[38] Appendix Figure 12 shows the distribution of markups and marginal costs.

Table 4: Quality's Insurance and Investment Costs

	Panel A: Insurance Cost ($\boldsymbol{\theta}_q^c$)		**Panel B: Investment Cost ($\boldsymbol{\mu}_k$)**	
Access	31.160	(16.690)	15.620**	(5.965)
Intermediate	108.400***	(12.800)	19.530***	(4.963)
Outcome	16.810***	(3.832)	15.000*	(6.516)
Patient	-244.300***	(57.540)	14.730*	(7.424)
Process	-175.600***	(27.560)	1.106	(4.718)
N	28966		5281	
R2	0.531		0.261	

Notes: This table reports the estimates of $\boldsymbol{\theta}_q^c$ in the marginal cost equation (9), and $\boldsymbol{\mu}_k$ in the investment cost equation (11). Values on the left are in dollars per member-month, while on the right are in millions per contract-year. Standard errors in parenthesis are heteroskedasticity robust. For further details on the marginal cost see Section 6.2.1. For further details on the investment cost Section 6.2.3. *p<0.05, **p<0.01, ***p<0.001.

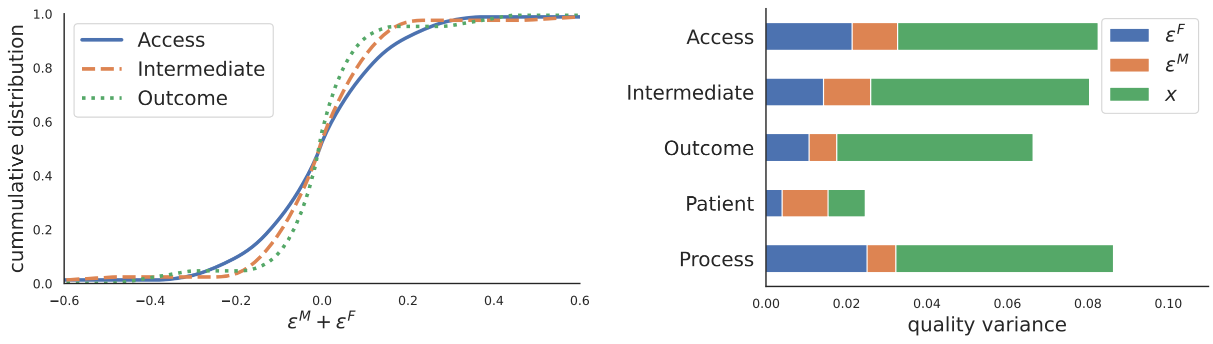
6.2.2 ***Quality shocks:*** Identifying quality investment costs is challenging because only their noisy outcomes are in the data. To address this challenge, I first show that the distribution mapping firm investments into quality ($F(\boldsymbol{q}|\boldsymbol{x})$) is identified and can be consistently estimated.

The intuition behind this result is the following. Consider two firms that offer products in two markets. The quality provided by each product is a combination of secretive network arrangements distorted by population shocks and firm-level shocks. As firms form beliefs about rival investments using market observables, controlling for those also controls for the joint distribution of investments. Consequently, any residual correlation in quality across firms within a market is driven by local population shocks. Conversely, any correlation across markets within a firm is due to firm-level shocks. Appendix Theorem 4 uses results from the non-parametric measurement error literature (Schennach, 2016) to transform this intuition into a conditional deconvolution result.[39] The proof delivers an analytic expression for the shock distribution in terms of the observable distribution of quality conditional on the set of market observables used by firms to form beliefs.

In the set of observables, I include regulatory features affecting firms' additional revenue, additional bundled services, plan types (e.g., HMO, PPO), and market sizes, as well as means, variances, and correlations between the same set of variables for rivals. Additionally, I include indicators for the presence of the top ten firms (by all-time enrollment) in the market. Overall, the vector contains over a hundred attributes observable by firms when investing. However, firms likely use only a few of these to form beliefs as rivals' investments are only relevant insofar they affect demand.[40] This observation suggests a sparse relationship between quality and the

[39] Similar results have been used in the auction heterogeneity literature (Krasnokutskaya, 2011).

[40] For example, knowing that one will be competing against Humana, who systematically commands a significant



(a) Shock distributions

(b) Variance decomposition

Figure 4: Quality shock distribution

Notes: These figures display the estimated distribution of quality shocks. Figure (a) shows the cumulative distribution function of the sum of shocks for each category. Patient and Process categories have been excluded for clarity. Panel (b) shows the fraction of the variance of the observed quality that is attributed to the quality target and the two types of shocks. Additional plots are provided in Appendix Section 6.3.1.

conditioning set. I leverage this by using the high-dimensional conditional density estimator of Izbicki and Lee (2017).[41] Finally, the estimator requires choosing the function that converts investments plus shocks to quality in every category k. Given that investments and errors are devoid of scale, the choice of $\Phi_k(\cdot)$ is a free normalization which I take to be $\Phi_k(x) = \Phi(x)(1-\underline{q}_x)+\underline{q}_k$ where $\Phi(\cdot)$ is the standard normal CDF, and $\underline{q}_k$ is the minimum value of quality k that firms can produce.[42]

In total, this estimation recovers the ten distributions modeled: two shock types for five categories, illustrated in Figure 4. The results show that shocks account for 39.6% of the variation in observed quality. Unsurprisingly, patients' assessments are the noisiest, as insurers can not contract for better reviews. Consistent with this, most of the variance in this category is due to market-level shocks. Firm-level shocks are most important in Process measures, as they are insurer-labor intensive, involving following up with patients and helping them schedule appointments for testing and care.

6.2.3 *Investment costs:* I estimate firms' investment cost parameters ($\boldsymbol{\mu}_{ft}$) by combining the optimality conditions associated with the investment problem and the estimated distribution of quality shocks. Appendix Proposition 3 shows that the marginal investment cost evaluated at the realized quality equates the expected marginal insurance profits, conditional on the realized

market share, is probably enough to render the attributes related to all other smaller rivals irrelevant.

[41] This estimator does not build on previous estimate. Because of this, and to offset the slower convergence rate of this class of non-parametric estimators (Horowitz and Markatou, 1996), I use the full 2009-2019 data for this estimation.

[42] While the domain of quality is the unit interval, in practice there are minimum standards. For example, an insurer can not contract with a hospital to act in a way that would actively harm patients.

quality, up to an error. Thus, the investment model satisfies the following regression:

$$\underbrace{\dot{\pi}_{jkt}}_{\substack{\text{conditional expteced}\\ \text{marginal insurance profits}}} = \underbrace{\frac{\partial I(\Phi_k^{-1}(\boldsymbol{q}_f) + \boldsymbol{\epsilon}_f^M + \boldsymbol{\epsilon}_f^F, \boldsymbol{\mu}_f)}{\partial x_{jkt}}}_{\text{marginal investment cost}} + \nu_{jkt} \qquad \mathbb{E}[\nu_{jkt}|\boldsymbol{q}_f] = 0 \tag{10}$$

Where $\dot{\pi}_{jkt}$ is the expected marginal profit of the third-stage, given the observation of $\boldsymbol{q}_f$. The key of this proposition is that the $\dot{\pi}_{jkt}$ is well-defined and not a function of the optimal investments, which are unknown. The expression for this variable is convoluted due to the non-differentiability of the scoring system and is therefore removed to the appendix. Its computational evaluation is also challenging due to the many degrees of integration. I tackle these challenges by leveraging techniques from the Quasi-Monte Carlo literature (Kuo and Nuyens, 2016) and approximating beliefs using a sufficient-statistic approach.

I estimate the marginal investment cost using a simple parametric specification. First, to avoid mixing contracts with very distinct cost structures, I limit attention to HMO and PPO contracts, which account for 81% of enrollment. This restriction excludes Private Fee-For-Service contracts, which do not form networks directly, and Regional PPO contracts, which have broad networks that often cross multiple state lines. Second, I separate the problem across states for each firm. While 18% of beneficiaries choose contracts offered in multiple states, the median multi-state contract has 80% of its population in a single state. Finally, for the included firms, omitting the implicit state index, I assume that the investment costs are additive across categories and expressed as

$$I(\boldsymbol{x}_{ft}, \boldsymbol{\mu}_{ft}) = \sum_{j \in \mathcal{J}_f} \sum_k \left(\mu_k (x_{jkt} - \underline{x}_{kt})^2 + \mu_{fkt}^F (x_{jkt} - \underline{x}_{kt})) \right)$$

Where $\underline{x}_{kt}$ is the state-category baseline quality, representing the lowest level of quality a firm can provide in a state. Anything above this level requires an investment in either forming a network or writing contracts that incentivize quality. Denoting the normalized observed quality, $y_{jkt} \equiv \Phi_k^{-1}(q_{jkt})$, this specification leads to the regression equation:

$$\dot{\pi}_{jkt} = 2\mu_k (y_{jkt} - \underline{y}_{kt}) + \mu_{f(j)kt}^F + \tilde{\nu}_{jkt} \tag{11}$$

The residual component, $\tilde{\nu}_{jkt}$, is the sum of two errors. First, it contains ν_{jkt}, introduced by substituting marginal profits with their expected values conditional on the realized quality. Second, it contains the error added by replacing the baseline qualities $\underline{x}_{kt}$, with the state-category-year minimum, $\underline{y}_{kt}$. Assuming that this second error is mean-zero conditional on y_{jkt}, equation (11) is a linear regression. Hence, I estimate μ_k and $\mu_{f(j)kt}^F$ using OLS, with firm-year-category fixed-effects to account for the cost-types.

Panel B of Table 4 displays the estimated coefficients. It shows that Intermediate Outcome quality is the most expensive to improve, while process measures are the cheapest. In comparative terms, the estimates suggest that contracts in the 75th percentile of Access quality invested an average of 5.16 million dollars more than contracts in the 25th percentile. Put together, the median quality firm invests 24.6% of its profits.

7 Scoring Design

The previous sections documented that scores shift demand and investment choices. They have also specified a model for the market and recovered estimates of the primitives governing product choice and quality production. Despite its simplifications, the model delivers reasonable estimates. Demand elasticities imply markups that agree with external evidence, and investment spending is moderate but meaningful. The key simplifications of static investments and homogeneous quality preferences are unlikely to significantly bias the subsequent analysis. Quality is a vertical attribute that varies rapidly in MA, and insurers in this market are likely to have already made their most significant long-term investments. This section leverages the findings of the previous sections and the assumption that consumers understand the scoring design (informed choice) to specify and partially solve the designer's problem. The following section studies scoring design without informed choice.

I assume that the designer seeks to maximize the expected weighted total welfare given by:

$$\max_{\psi\in\Psi}\int\left(\underbrace{CS(\psi,\boldsymbol{q})}_{\substack{\text{Consumer}\\ \text{surplus}}}+\rho^F\underbrace{\sum_f V_f(\psi,\boldsymbol{q})-I(\boldsymbol{x}^*_f(\psi),\mu_f)}_{\substack{\text{Insurer}\\ \text{profit}}}-\underbrace{\rho^G Gov(\psi,\boldsymbol{q})}_{\substack{\text{Government}\\ \text{spending}}}\right)dF(\boldsymbol{q}|\boldsymbol{x}^*(\psi)) \tag{12}$$

The objective sums consumer surplus and insurer profits and subtracts what the government spends on subsidizing enrollees in MA relative to the cost of insuring them under TM. To do so, the designer publicly announces and commits to a deterministic scoring rule (ψ) that partitions the space of quality into distinct ratings, assigning higher ratings to better quality. I focus on this class of deterministic monotone designs as they are common, incorporate the MA Star Rating system, and are optimal under certain scenarios.[43]

[43] All deterministic quality certifications fall within this class (crash test, organic labels, high-in-sugar food labels, etc.). See Dworczak and Martini (2019) for a proof of optimality for similarly defined scores under exogenous quality. Their definition allows some segments to be revealing, which I explore in the appendix.

The key tension behind the designer's choice is a trade-off between eliminating information frictions and regulating quality. Intuitively, if quality were exogenous, full information would be optimal for the designer, and the solution would consist of finding the score that best approximates the full-information outcome. However, when quality responds to scoring, the optimal design addresses the Spencian distortion caused by firms' market power over quality, resulting in a coarse scoring system limiting consumers' information. In this problem, the Spencian distortion states that the investment vector which maximizes the designer's objective differs from the one firms would optimally choose under a fully informative score. The tension caused by this distortion is apparent in the objective. Ignoring government spending and taking $\rho^F = 1$, full information maximizes the integrand but forgoes any attempt to regulate investments ($\boldsymbol{x}^*(\psi)$) and thus the distribution of quality in the market ($F(\boldsymbol{q}|\boldsymbol{x}^*(\psi))$).

The same tension between information and regulation also appears within the consumer surplus component. As the informed choice assumption allows consumers to hold biased beliefs about the quality distribution, the regulator evaluates the true consumer surplus (Train, 2015), given by :

$$CS(\psi, \boldsymbol{q}) = \underbrace{CS_0(\psi, \boldsymbol{q}) + \zeta}_{\substack{\text{Consumers'} \\ \text{expected surplus}}} + \underbrace{\sum_{j \in \mathcal{J}} \gamma'(\boldsymbol{q}_j - \mathcal{E}[\boldsymbol{q}_j|\psi(\boldsymbol{q}_j)])(\int \frac{s_{ij}(\psi(\boldsymbol{q}))}{|\alpha_i|} di)}_{\text{Ex-post error correction}} \tag{13}$$

The first term is the well-known ex-ante logit consumer surplus, with ζ its unidentified location (Small and Rosen, 1981). The second term adjusts the surplus to account for the gap between consumers' beliefs about the product's quality and its actual value.[44] As the designer evaluates this term in expectation, it captures the accuracy of consumers' beliefs relative to the true distribution of quality. That is, the designer is concerned with ex-post mistakes by consumers, given her ex-ante knowledge of optimal investments. In turn, these investments vary with the design as it alters product demand conditional on quality.

Solving the designer's problem is challenging. Scores are discontinuous functions from multiple dimensions to a scalar, making typical approaches to functional optimization unusable. The approach I develop and describe in detail in Appendix 7.3 relies on the intuition of Kamenica and Gentzkow (2011) that choosing signals is akin to choosing a distribution over posterior beliefs. In the case of scoring design, the analogous statement is that each score generates a distribution over qualities, rating valuations ($\mathcal{E}[\boldsymbol{q}|r, \psi]$), and marginal quality costs ($\boldsymbol{\theta}^{c\prime}\boldsymbol{q}$). This observation enables a "gridding" strategy, which first evaluates the objective over a large collection of values covering

[44] In this expression, $\tilde{D}_j$ is a weighted sum of the demand for plan j. Consumers whose valuation of quality is higher relative to premiums are given a larger weight.

the space of qualities, rating valuations, and costs and then associates each score with a distribution over the grid.[45] The final key to solving the problem is a simple proposition (Appendix Proposition 4) that shows that every monotone partitional score is a composition of a *reduction* function that maps multiple dimensions down to a scalar, and a *cutoff* function that segments the scalar to scores. Moreover, the proposition shows that a finite order polynomial approximates the reduction function arbitrarily well.

The remainder of this section is organized as follows. First, I present what I call the *best linear substitute* for the MA Star Ratings, an alternative design that, like MA's, uses nine scores and a linear function to reduce multiple dimensions of quality into one. I show that this design can improve welfare, and I unpack the margins through which this improvement occurs. Next, I compare this design to a full-information counterfactual, which delivers two key results: the optimal design for MA is necessarily coarse, and there is a Spencian distortion in the market. To understand which aspects of the new design are fundamental for welfare improvements, I study the design of quality certifications. This analysis delivers two additional results: simple designs can be better, and welfare is non-monotonic in the informativeness of a score. Finally, I study how strategic manipulation of the score by the regulator towards her private preferences for quality results in a welfare loss and an ineffective scoring system. I delegate additional results concerning more sophisticated designs to the appendix; their improvement over the best linear substitute is minimal. The appendix also provides results on the effects of competition and geographic aggregation on scoring design. All results are for a subset of markets in 2015, covering nearly 22 million beneficiaries.[46] I show results for the case of $\rho^F = 1$ and $\rho^G = 0$, and present those for consumer surplus maximizing designs ($\rho^F = \rho^G = 0$), and the full objective ($\rho^F = \rho^G = 1$) in the appendix.

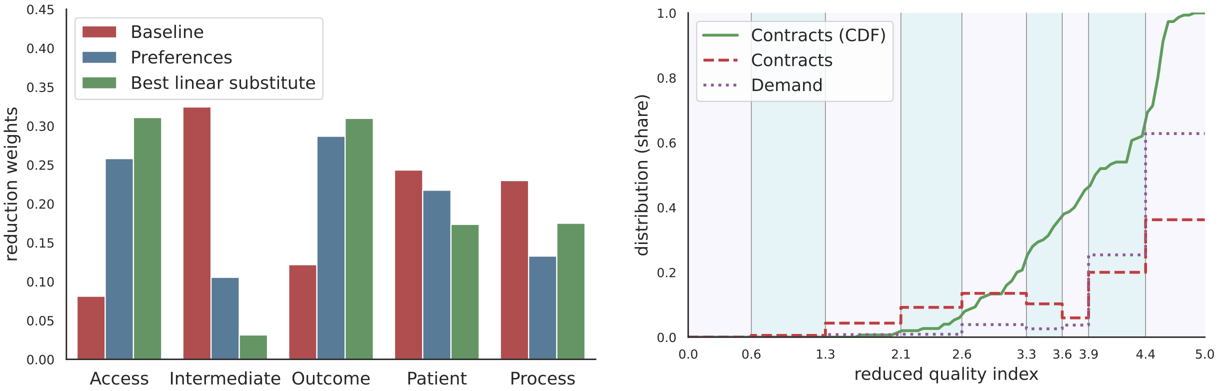
7.1 Best linear substitute

The best linear substitute design maximizes total welfare using the same technology as the MA Star Ratings. Like MA's system, this alternative design classifies quality into nine distinct scores and uses a linear mapping (e.g., a weighted average) to aggregate quality dimensions. To find the best linear substitute, I optimize the linear mapping that reduces multidimensional quality to an index (i.e., the reduction function) and the set of cutoffs that partition the index into scores (i.e., the cutoff function). Figure 5 displays the resulting design.

The first noticeable difference between this new design and the Star Ratings is a shift in the weights used to combine dimensions. Relative to the baseline, the new design removes weight from

[45] These spaces are compact given that the space of quality is compact.

[46] The subset is given by the set of counties covered by the MCBS after the HMO/PPO restriction.



(a) Reduction weights

(b) Quality distribution and cutoffs

Figure 5: Best Linear Substitute Design

Notes: These figures display the design of the best linear substitute scoring system. The figure on the right shows the relative weight placed on each dimension of category, compared to CMS's weighting scheme and consumers' preferences. As the reduction function of this design is linear, it is fully represented by these five values. The figure on the left shows the cutoff placement for the the nine scores used by this design. The green solid line shows the cumulative distribution of contracts over the index quality, while the dashed red one shows the expected share of contracts attaining each score. I expand the scale of the reduced index to match that of the Star Ratings.

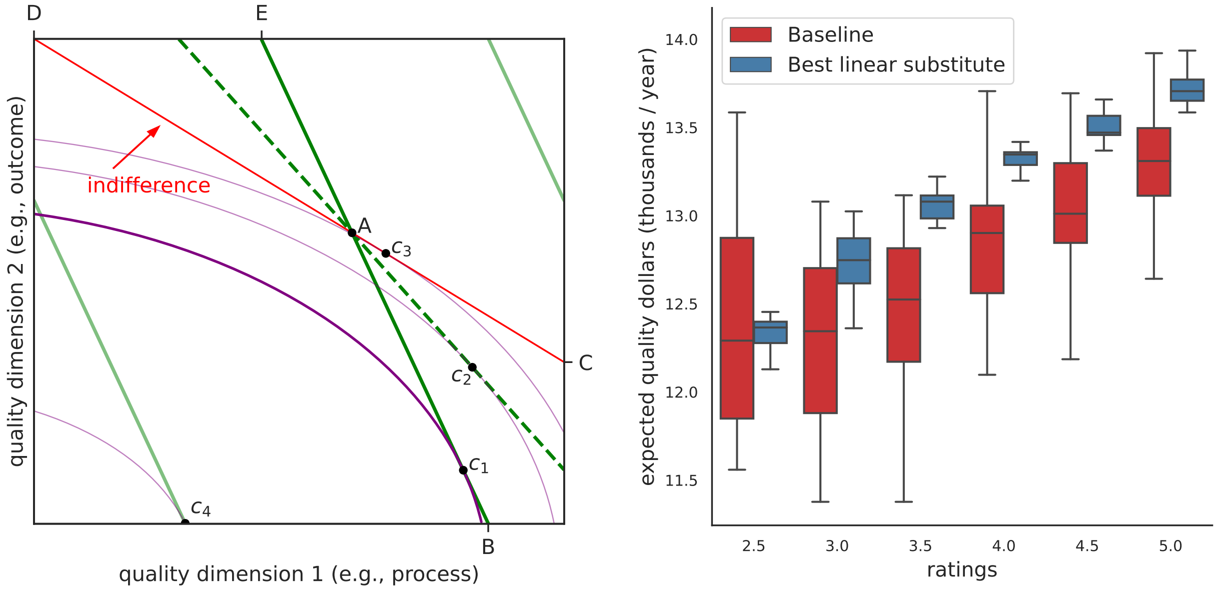
the Intermediate Outcome category and shifts it towards the Access and Outcome categories. This rearrangement of weights is in the direction of consumer preferences, improving the alignment between the rating thresholds and consumers' iso-utility curves. This modification plays both a regulatory and informational role.

On the regulation side, reallocating reduction weights is a way to control how firms distribute their overall quality investments across categories. Like every coarse scoring system, the Star Ratings hide from consumers both the product's level and composition of quality within a scoring bin. This coarseness of information creates a distortion illustrated in figure 6.a in a scenario of two quality dimensions, four scores, and no investment risk. The figure indicates that a firm choosing to receive a score of three will optimally invest at a point on the frontier of the rating interval that is tangent to its iso-cost curve (point c_1). Thus, once a firm has decided which score it wishes to receive, its relative quality investment is only a function of the rating frontier and its cost curve, disregarding consumers' preferences. For the designer, this incentives distortion produces a multitasking moral hazard problem as in Holmstrom and Milgrom (1991). The designer would like firms to invest more in categories that consumers value more, while firms have incentives to invest only according to their cost. The best linear substitute design addresses this problem by improving the alignment of rating thresholds and consumers' iso-utilities. Therefore, any substitution of quality categories along the frontier leaves consumers nearly indifferent. The optimal design does not align the two curves perfectly because some firms face onerous convex costs of improving certain categories. A perfectly aligned threshold would make some of these firms change their desired scoring level, creating a discontinuous drop in the overall quality they provide, lowering welfare.

Aligning the scoring threshold with consumers' quality preferences also alters the informational content of the scores. This happens through two channels. First, an increase in the contribution of a category to the score augments the correlation between a product's category-level quality and its rating. Therefore, the best linear substitute design renders the score more informative of variation in Outcome quality at the expense of less information about Intermediate Outcome quality. The second channel is through a decrease in the probability of ex-post mistakes in insurance choice. Figure 6.a illustrates that the misalignment of rating thresholds and iso-utilities creates regions of *misclassification*. Plans falling in the triangle *DEA*, receive a lower rating than those in the triangle *ABC*, but consumers prefer the former over the latter. Narrowing the regions of misclassification has a non-trivial effect on this market. Figure 6.b shows that the MA Star Ratings have a substantial overlap in expected quality utility between ratings. This overlap makes it harder for consumers to choose among products and makes the rating less useful for them, in turn eroding the score's ability to marshal demand and thus affect quality.

The second design choice is the cutoff placement. Figure 5.b shows that the new design has coarser bins in the extremes and thinner in the center-right. Contrasting the bins with the resulting distribution of contracts, the cutoffs pool together a vast portion of low-quality contracts and are more precise about qualities occupying the 3.3 to 3.9 index. Like the MA ratings, the new design has two scores located in the lower spectrum of quality where, in expectation, no product lands. These scores are used when products receive large adverse quality shocks, and dividing this spectrum into different scores dampens firms' losses under those shocks. On the other extreme, the design allocates a large bin to the maximum quality. However, the maximum quality in the market is near the bin's lower boundary, and consumers have primarily accurate beliefs about this distributional feature, resulting in a precise, high-quality signal. This signal contrasts with the one consumers receive for quality in the second-to-last bin. These two last scores create a similar effect as in the theoretical analysis of Section 2: the lower signal pools low qualities and receives a low demand; the higher signal accurately reveals high-quality, leading to higher demand. This feature explains why the new ratings result in over 40% of contracts being top-rated, serving 60% of consumers.

The estimated welfare gains from replacing the Star Ratings with the best linear substitute design are shown in Table 5. Per Medicare consumer-year, the alternative design increases surplus by \$131.6, and firm profits by \$518.7, while slightly decreasing government spending. The consumer surplus gain is equivalent to 2.1 months of average premiums (part C + D) in the baseline, and the additional profits correspond to over a twofold increase for firms. This increase is due to



(a) Allignment illustration

(b) Quality overlap per rating

Figure 6: Effects of aligning category weights with consumers' preferences

Notes: Figure (a) shows the theory with two quality dimensions and four scores, each delimited by the solid green lines. The line *EB* is the second scoring cutoff, and *DC* is consumers' iso-utility. Products above *EB* obtain a score of three, and below it, two. The misclassification region is the sum of polygons *DEA* and *ABC*. Products in *DEA* get two stars, but consumers prefer them over those in *ABC*, which get three stars. The new design corresponds to the green dashed line, which is tilted towards the iso-utility, narrowing the misclassification region. The optimal misclassification is not zero because it might detract firms from investing. The purple concave lines represent a firm's iso-cost curve, such that in the baseline, it chooses the quality combination point c_1 and gets three stars. In the new design, it chooses c_2. However, perfect rating alignment leads the firm to choose c_4 instead of c_3 because the gain in demand is lower than the increase in cost. Figure (b) shows the ranges of quality utilities binned in each score in the baseline and the best linear substitute design. The baseline has extensive overlap between ratings, implying large areas of misclassification. The best linear substitute leads to negligible misclassification despite being imperfectly aligned.

63% higher profits per MA enrollee and a significant market expansion as consumers switch from TM to MA. Higher quality and better information about it cause this expansion effect. I interpret this finding as indicating that a large fraction of consumers choose TM because the narrow MA networks have uncertain quality. In the counterfactual, once quality improves, many consumers substitute, often to plans that cost less to subsidize than TM, leading to a decrease in spending of 0.8% per Medicare beneficiary.

To unpack the channels through which the welfare gains occur, I gradually incorporate the counterfactual equilibrium's information, quality, and prices. Table 6 shows the result when evaluating the designer's objective at the expected equilibrium outcome. As shown, introducing the new scores while keeping quality and prices at their baseline levels delivers a large fraction of the gains. Consumer surplus slightly exceeds its final equilibrium value, insurers obtain about

Table 5: Welfare changes in redesigned system under informed choice

	Linear	**Full**	**Certification**		
	Substitute	**Information**	Best linear	CMS	Preferences
Δ Consumer Surplus	131.6	136.0	151.8	68.6	122.8
Δ Firm Profits	518.7	488.6	480.5	169.1	486.5
Δ Gov. Spending	-74.1	-69.1	-88.4	-85.3	-94.0
Δ Total Welfare	650.3	624.6	632.4	237.7	609.2
MA share	56.9%	54.4%	56.0%	41.3%	54.9%

Notes: This table displays the welfare effect of redesigning the scoring system, relative to the MA Star Rating baseline. All values are in 2015 dollars per Medicare beneficiary. Government spending corresponds to the change in subsidy and rebate payments, including the cost of subsidizing TM (FFS costs). The baseline simulated market share of MA is 27.8

60% of their final gains, and the overall demand for MA more than doubles. Allowing quality to change nearly doubles consumers' surplus, moderately increases firm profit, and expands MA further. Finally, incorporating the equilibrium prices leads to increased premiums and a slight erosion of cost-sharing benefits. These changes create a substantial transfer of welfare between consumers and firms, with profits increasing by 75% and surplus dropping by over 50%. Overall, the analysis shows that disclosure's informational and regulatory powers have a near-identical impact on consumer surplus, and prices are a mechanism through which the market redistributes the welfare gains across agents.

7.2 The full information benchmark

Comparing the previous results against the full information benchmark reveals whether the market suffers from a Spencian distortion and sheds light on the optimal score for this market. First, suppose quality is higher under the best linear substitute design. In that case, it must be that the regulator would prefer quality to be higher than what it achieves under full information. Second, if the total welfare under the alternative design is higher than under the full information benchmark, then it must be that the optimal design for MA is coarse. Crucially, this would allow me to state something about the optimal design without knowing which constraints the regulator faces. For example, technological and behavioral considerations might constrain the designer to weighted-average systems (i.e., linear reductions) with at most nine scores, in which case the best linear substitute is optimal, or it could be that designer is unrestricted in the number of scores. The best linear substitute design is feasible construction and finding that it dominates full information

Table 6: Welfare decomposition for alternative designs

	Δ Firm Profits	Δ Consumer Surplus	MA Share	Avg. Premium	Avg. Benefits
Best linear substitue					
+Δ Information	316.3	239.1	64.8%	22.4	74.3
+Δ Quality	334.3	444.4	72.0%	21.8	75.7
+Δ Prices	511.4	222.7	67.3%	41.2	72.4
Certification					
+Δ Information	83.7	-9.5	44.5%	23.1	68.3
+Δ Quality	267.1	389.2	67.5%	20.9	76.1
+Δ Prices	437.9	178.0	62.7%	40.9	72.3

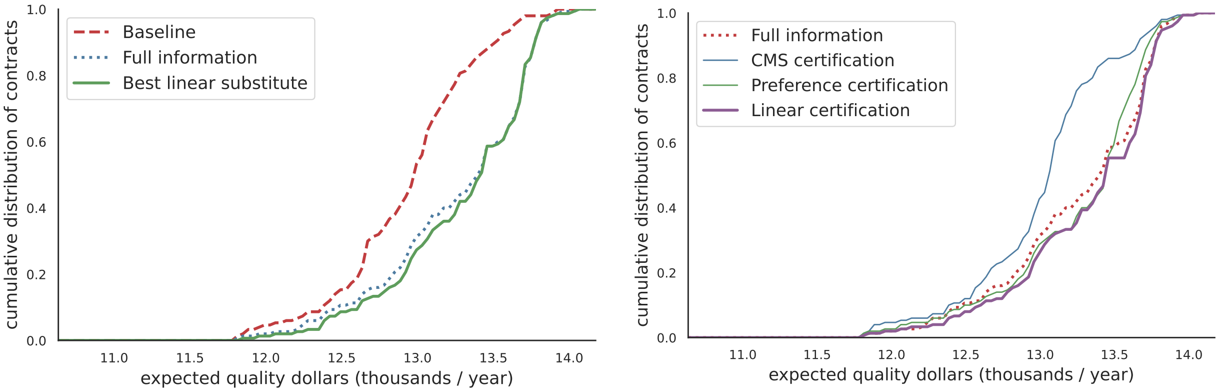
Notes: This table displays the welfare change from gradually incorporating the equilibrium outcome of the best linear substitute and the quality certification designs. To avoid confounding the effect of investment risk, this decomposition is done at the expected realized quality given insurers' optimal investments. Because the welfare components are locally concave to the equilibrium, the values presented in this table are slightly larger than those obtained for the overall welfare. All values are in 2015 dollars per Medicare beneficiary.

would prove that the unknown optimal design is also necessarily coarse.[47] Whether the designer's choice set includes any fully informative design is unclear, making this alternative a theoretical benchmark rather than a feasible solution.

The second column of Table 5 shows the welfare change under the full information benchmark. Compared to the best linear substitute, consumers are slightly better off, but total welfare is lower. This finding validates that the total-welfare optimal scoring design for MA is coarse. Appendix 7 shows that the same finding holds if we consider the designer as maximizing only consumer surplus. Figure 7.a shows that the quality distribution under the best linear substitute design dominates the full information distribution in the first-order stochastic sense. This finding is evidence that MA insurers would continue to underprovide quality even under full information due to their market power over quality as illustrated in Section 2.

Finding that the optimal design is necessarily coarse might have important policy implications. In particular, it shows a slight contradiction between CMS's stated goals of informing consumers and promoting quality (Medicare Payment Advisory Commission, 2013). The results indicate that when market power over quality exists, providing consumers with detailed quality information might be worse for them and reduce insurers' incentives to invest in quality. The minor welfare differences between the best linear substitute and the full-information benchmark falsely

[47] Appendix 7 shows optimal designs under more sophisticated reductions and scoring restrictions.



(a) Best linear substitute

(b) Quality certification

Figure 7: Quality Distribution

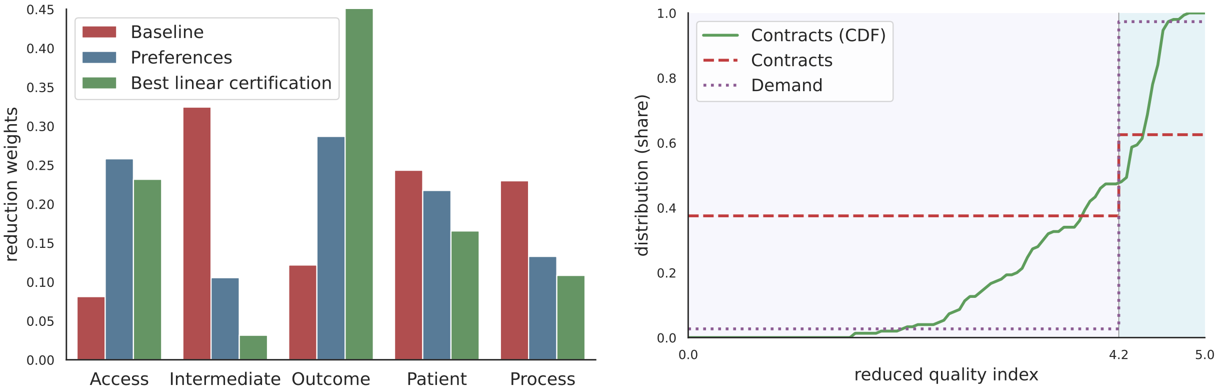
Notes: These figures display the cumulative distribution of contracts relative to their expected quality. The figure on the left presents the distributions compared in the Best Linear Substitute design, while the figure on the left shows the ones for the certification designs.

suggest that maximizing the informativeness of scores might be a good heuristic. The following section will show that this is far from true, as the welfare gains from scoring do not increase with informativeness.

7.3 Optimal certification

Quality certifications are the most common type of quality disclosure rule. Beyond their simplicity, their prevalence might be due to their ability to affect quality. In particular, certification pools all uncertified low qualities into one score and all high qualities into another. Thus, products below the certification threshold receive the worst possible signal any monotone deterministic score could provide, while those above it obtain the best signal. Intuitively, this will tend to make consumers buy fewer uncertified products, creating incentives for firms to invest in quality. If these incentives are strong enough, few firms will offer uncertified products, and those certified would offer qualities near the certification cutoff to maximize profits. This is precisely the mechanism behind the monopoly regulation example of Section 2, which results in consumers buying a product with exact beliefs about its quality despite obtaining only a dichotomous signal of quality.

To test whether this intuition applies to an empirical context, I solve for the optimal quality certification for MA. I focus on designs that, like the baseline, use a linear function to aggregate quality dimensions and present designs using higher-order reductions in the appendix. Figure 8 shows the resulting weights and the cutoff placement. As in the previous solution, this design improves the alignment of reduction weights and consumers' preferences. Also, the certification



(a) Reduction weights

(b) Quality distribution and cutoffs

Figure 8: Best linear certification design

Notes: These figures display the design of the best linear certification scoring system. For further details on the plotted components see the details of figure 5.

cutoff is placed in an index that has a similar quality-utility value as the threshold to be top-rated in the best linear substitute.[48] The share of products choosing to receive certification is 62.1%, serving 97.3% of consumers, who in turn obtain a mean ex-post quality utility of $16 per month.

The welfare gains from the optimal placement of cutoffs and weights under certification are shown in the third column of table 5. Both firms and consumers prefer simple certification over the more sophisticated status-quo or the full information benchmark. This last finding confirms the intuition of the stylized theoretical framework that the certification can be used to offset firms' market power over quality. Furthermore, Figure 7.b shows that the distribution of quality under certification dominates that of full information, confirming that this design can narrow the Spencian distortion gap. A corollary of these results is that total welfare is non-monotonic in the score's informativeness. However, this is only true because of the endogeneity of quality to the score. Table 6 shows the decomposition of welfare, evaluated the expected quality outcomes. The table shows that replacing the Star Rating with certification while keeping quality and prices fixed would harm consumers. The gains from certification stem exclusively from the quality effect.

The simple structure of quality certification makes it easier to examine the effect of reduction weights on total welfare. To study this, I solve for the optimal certification while holding CMS's relative weighting scheme fixed and using consumers' preferences to weight categories. That is, the CMS-based certification shows the welfare gains that the market could attain by transitioning to an optimized certification design keeping CMS's priorities over quality as they currently are.

[48] The mean posterior belief for certified products is $16.1 dollars per month, while for top-rated products under the best linear substitute design it is $16.27.

Instead, the preference-based certification shows the welfare gains if the designer fully aligned weights with preferences such that the system would guarantee that a higher rating implies a higher quality utility. The computed welfare effects are shown in the last two columns of table 5, and the cutoff placements are in the appendix. Figure 7.b contrasts the distribution of quality under the three types of certification.

This exercise indicates that CMS could attain a similar quality distribution and welfare as in the status-quo using quality certification. This finding is consistent with anecdotal evidence that insurers in MA aim for four stars, suggesting a behavior similar to the one excepted under certification.[49] Under this design, 52.1% of contracts obtain a certification, 91.7% of consumers purchase a certified product and obtain an ex-post average quality utility from certification of $15.4 per month. Aligning the weights with consumers' preferences significantly shifts quality production and demand. 62% obtain certification serving 95.8% of consumers, who now obtain an average ex-post quality of $15.9. Replacing the weights with consumers' preferences delivers almost the entirety of the welfare gain of the more sophisticated design. In general, the results of this work indicate that consumers' willingness to substitute one quality dimension for another serves as an adequate rule-of-thumb replacement for a weighting scheme.

Overall, the results of this section have a clear connection with evidence from other markets and policies in MA. The literature has documented significant supply responses to quality certification in settings such as appliances (Houde, 2018a) and groceries (Barahona et al., 2020). My results suggest that similar incentive schemes can be used to promote quality even in scenarios with imperfect control (i.e., risky investment or effort) and multidimensional quality. It also shows the potential benefit of aligning scoring incentives with consumer preferences in settings where multitasking has been a concern, such as nursing homes (Feng Lu, 2012). Finally, the results highlight the importance of creating large incentives gaps in order to promote quality. Surprisingly, an influential advisory commission recently suggested that Congress remove such incentives – or what they call "cliff effects" – from other quality promoting programs in MA (MedPAC, 2020). This section indicates that cliff effects are crucial to promoting quality and might also improve plan quality information.

7.4 Strategic score manipulation

A caveat of the previous analysis is that the designer might value quality beyond its impact on total welfare or subsidy spending. As CMS bears the brunt of the cost if consumers' health deteriorates, the designer and consumers might disagree about the value of reducing long-term health risks. This

[49] I learned this from interviewing the director of Star Rating Analytics at one of the largest insurance companies in MA.

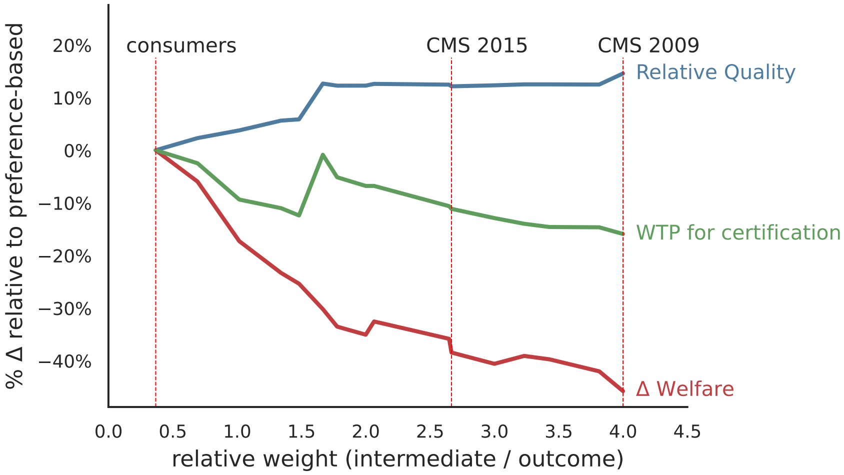


Figure 9: Welfare and certification value under manipulated score

Notes: This figure displays how investment, certification value, and welfare change as the score's weight on Intermediate quality relative to Outcome quality increases. The discontinuities are due to discrete changes in the number of firms choosing to certify, leading to abrupt changes in investments.

conjecture could explain why the baseline weighting scheme differs from consumer preferences. Nevertheless, it is unclear whether there are gains from nudging consumers away from their preferred choices. If the regulator distorts the score to highlight specific products, consumers might respond less to the signal, reducing supply incentives and potentially offsetting the gains from the nudge.

To evaluate this possibility, I solve for different certification equilibria starting from the preference-based design and adjusting the weights of the Intermediate and Outcome categories. I choose the weights to span CMS's designs between 2009 and 2019, keeping the overall contribution of the two categories fixed. The results, shown in Figure 9, indicate that increasing the relative contribution of Intermediate relative to Outcome increases the relative quality investment produced and purchased. However, the change in quality plateaus rapidly at around 12.5%.[50] The reason is that consumers' willingness to pay for certified products deteriorates rapidly as the signal becomes less representative of the information they seek.[51]

Overall, the results suggest that scores are not practical mechanisms for consumer nudging. To justify the observed weighting schemes, the regulator would have to value a small reallocation of

[50] These percentages are relative to the preference-based certification design. This certification is fast to compute and makes it easier to illustrate the effect of distorting the alignment. Thus, the ratio of Intermediate to Outcome quality of the average chosen product increased by 13% relative to the same ratio under the preference-based certification equilibrium.

[51] The level of certified quality (not shown) also decreases because of dampened demand incentives, dropping from $16 to $15.5 dollars per month in ex-post utility. Uncertified quality remains stable at $13.4.

quality across categories by hundreds of millions of dollars. As the reallocation implied by CMS's weights is away from Outcomes, it is difficult to assess whether meaningful quality-of-life effects of improving Intermediate quality justify such a distortion.

8 Robust scoring design

In this final section, I explore how to design scores when consumers' preferences for quality are unknown. I consider the case in which the designer only knows that preferences are within a particular set denoted Γ. As appendix theorem 1 shows, this would be the case if consumers did not understand the scoring design changes in the MA market, and instead had exogenous and unchanging beliefs about what a star rating means. In this scenario, a designer lacking alternative sources of elicited preferences and beliefs, would be unable to identify consumers' beliefs about quality but could derive bounds about their preferences for different dimensions. Knowing only the set Γ, I consider the design objective that maximizes total welfare under the worst-case preferences:[52]

$$\max_{\psi\in\Psi}\min_{\gamma\in\Gamma}\int\left(\underbrace{CS(\psi,\boldsymbol{q})}_{\substack{\text{Consumer}\\ \text{surplus}}}+\rho^F\underbrace{\sum_f V_f(\psi,\boldsymbol{q})-I(\boldsymbol{x}^*_f(\psi),\mu_f)}_{\substack{\text{Insurer}\\ \text{profit}}}-\underbrace{\rho^G Gov(\psi,\boldsymbol{q})}_{\substack{\text{Government}\\ \text{spending}}}\right)dF(\boldsymbol{q}|\boldsymbol{x}^*(\psi)) \tag{14}$$

As consumers' beliefs are unknown and independent of the design, the designer's toolkit is limited to using the nine pre-existing scores. Her objective is to use these scores to maximize welfare under the worst-case scenario of preferences. Importantly, this objective would remain the same if consumers had heterogeneous preferences for quality. This cautious approach matches the decisions of an imperfectly informed regulator that risks significant political or legal losses from implementing a new design that worsens outcomes. Appendix Proposition 5 shows that the interior minimization is equivalent to a linear equilibrium constraint, which facilitates solving this problem.

This exercise complements the work of the previous sections in three ways. First, it helps to disentangle the mechanisms by which the score affects the market. In particular, in the previous sections, the designs coordinated consumers by changing the assignment of scores to products and their beliefs about the quality represented by those scores. In this exercise, the second channel will

[52] Preferences γ are relevant only for consumer surplus as, conditional on the identified fixed valuation for ratings, the demand is independent of consumers' preferences for quality.

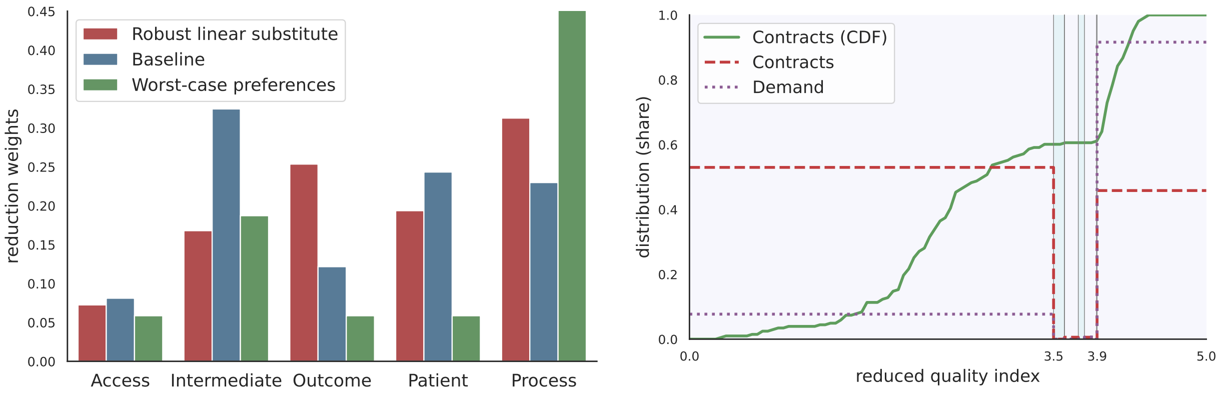
be shut off, showing that scoring design can be effective even if consumers are unaware of design changes. Second, it highlights the importance of creating transparent and well-communicated scoring systems. The gap between the results of this section and the previous – consumers' understanding of the design – appears fixable by an informational regulator. Finally, it provides an alternative solution for the cautious regulator (or reader) unnerved by the assumption of informed choice. Moreover, it allows me to compute a worst-case scenario for the previously proposed designs.

I solve for the linear reduction and associated cutoffs that optimize the robust design problem.[53] As in the main analysis, I consider two solutions, one using all nine scores and another using only one and five stars to create a certification scheme. The computed design for the *robust linear substitute* is shown in Figure 10 and the certification design is shown in the associated appendix section. Intuitively, given that consumers' preferences are adversarial, the design attempts to promote uniform quality production. The non-uniform bounds on preferences and firms' cost heterogeneity skew the optimal weight placement, resulting in a higher relevance for the Process, Patient Experience, and Outcome categories than for Access and Intermediate Outcomes. The cutoff structure of this design is peculiar, resulting in a "padded" certification scheme. As in standard certification, the system assigns one star to all low qualities and five stars to all high ones. The padding of this design is that it allocates five intermediate stars in the gap between the bottom and top rating. In expectation, no firm lands in this narrow gap. However, it does increase firm profit by reducing their investment risk when attempting to reach the top rating.

The fundamental mechanisms of the robust design are the same as with informed choice. Consumers penalize lower-rated products leading to an increase in investment and certification. In equilibrium, top-rated products in both robust designs deliver a quality that is very similar in value to those of five-star plans in the baseline. However, a larger share of plans obtain this rating and serve about 91% of consumers. Thus, the robust designs eliminate a range of intermediate qualities that distort firms' incentives without significantly benefiting consumers. The vertical nature of quality and the limited cost heterogeneity makes it welfare-enhancing to trim these intermediate qualities.

The welfare effect of the robust linear substitute and the certification designs are displayed in Table 7. The estimates indicate that both improve the worst-case scenario welfare. As the more flexible design results in a certification scheme, the welfare difference between the two solutions is minimal. In both cases, as consumers' preferences are unknown and adversarial, the consumer

[53] In order to discipline the worst-case preferences, I further restrict Γ such that γ are between half the lowest estimated value and twice the highest among all quality dimensions. This means that the highest quality product can be worth anywhere between \$4133 and \$44984 per year in premiums. Otherwise, the worst-case scenario often derives zero utility from the quality dimension with the highest investment, which is unreasonably harsh.



(a) Reduction weights

(b) Quality distribution and cutoffs

Figure 10: Robust certification design

Notes: These figures display the reduction weights, worst-case preferences, and cutoff placement of the robust linear substitute design. The figure on the left displays the reduction weights in the baseline and the new design. The worst-case preferences for this problem are also illustrated. The figure on the right shows the cutoff placements and the resulting distribution of contracts.

surplus gain is mechanically small. However, the results still indicate an expansion of MA and a reduction in public spending.

The robust design objective can also be used to evaluate the worst-case preference scenario for the designs derived under the assumption of informed choice. The last two columns of Table 7 display these effects for the best linear substitute and linear certification design. Both have a positive welfare impact, indicating that even if they were designed with wrong conjectures about consumer preferences for quality, they would still lead to overall improvements. This indicates that redesigning the system to jolt investment and competition is almost always beneficial, even if the relative importance of quality investments is erroneously measured. This finding, of course, is driven by quality being a vertical attribute of products: all else fixed, more quality is always better in every dimension.

Finally, the last column of Table 7 shows the worst-case scenario for the CMS-based certification design. This design uses CMS's weighting scheme to aggregate quality dimensions, one of the main differences between the robust and main certification designs. The results indicate that the CMS weighting scheme is particularly well suited to address the robust problem, although the cutoff placements could be improved. In particular, CMS's weights are nearly optimal within the class of linear reductions and result in a better worst-case scenario than both the main linear substitute and certification designs. This observation suggests that the CMS weighting decision might be driven by an abundance of caution about misrepresenting consumers' preferences.

Table 7: Worst-case welfare changes in redesigned system

	Robust		**Previous Designs**		
	Substitute	Certification	Substitute	Certification	CMS-based
Δ Consumer Surplus	5.6	12.7	-322.9	-249.4	72.1
Δ Firm Profits	288.6	277.0	428.7	415.1	173.0
Δ Gov. Spending	-154.6	-146.1	-163.2	-175.5	-107.9
Δ Total Welfare	294.3	289.7	105.7	165.7	245.1
MA share	50.7%	49.6%	60.0%	57.0%	42.0%

Notes: This table displays the welfare effect of redesigning the scoring system, relative to the MA Star Rating baseline under the worst-case consumer preferences for quality. The baseline for welfare comparison also evaluates the worst-case preferences given the Star Ratings. The Previous Designs columns evaluate the worst-case for the designs derived using the informed choice assumption. The CMS-based columns presents the quality certification design that uses CMS's weighting scheme. All values are in 2015 dollars per Medicare beneficiary. Government spending corresponds to the change in subsidy and rebate payments, including the cost of subsidizing TM (FFS costs).

9 Conclusion

I study the problem of designing a scoring system for firms with market power over quality. Using data from Medicare Advantage in 2009-2015, I show that scores shift demand across products and alter insurers' quality investments. Leveraging individual-level choices and variation in the scoring design, I estimate a structural model of demand and supply responses to scoring. I specify the problem of a welfare-maximizing designer and use the model estimates to evaluate alternative designs and find local optima. The analysis presents three novel findings.

First, I show that the optimal disclosure policy for MA involves coarsening quality information. Total welfare is neither increasing in how informative a scoring system is nor is it maximized at full information. The central mechanisms for this behavior are that quality responds to ratings and is underprovided by firms under full information. Scores can marshal demand to offset firms' market power over quality, increasing total welfare. I propose an alternative design that increases welfare substantially, with half of the improvement stemming from better quality information and the remainder from the endogenous quality responses. This finding highlights the importance of considering the effect of information policies on the endogenous supply of quality and evaluating how these might be coordinated or overlap with other efforts to alter the market. This result also provides empirical support to the growing theory on scoring design (Rodina and Farragut, 2016; Ball, 2019; Hopenhayn and Saeedi, 2019; Boleslavsky and Kim, 2018; Zapechelnyuk, 2020).

Second, I find that a well-designed quality certification can vastly improve welfare and tightly approximate the effect of more sophisticated scores. In MA, I find that certification achieves

97% of the welfare gains of an optimized nine-scores system. The results indicate that cliff-effects in firms' incentives are fundamental in promoting quality, contradicting some recent policy recommendations (MedPAC, 2020). This finding also highlights the contradiction between CMS's effort to inform consumers, promote quality, and improve overall welfare in the market. Instead, my results show that scoring designs that reveal very little information can result in more informed purchases of higher quality. Finally, these results also provide evidence on why some certification schemes have been exceptionally effective despite their simplicity (Barahona et al., 2020).

Third, I find that skewing the score's information away from what consumers care about quickly erodes its value, and thus its ability to alter the market's outcome. This finding has important implications for scoring designs in general, particularly for CMS's practice of seeking design feedback from the industry. The results suggest that CMS should instead elicit consumers' preferences for quality when designing its quality aggregation scheme. A combination of theoretical and empirical results in this paper also highlights the importance of clear communication and transparency in scoring design. They play an essential role in eliciting consumers' preferences and in the score's effectiveness as both an information and regulatory policy.

My results point the way to several possible extensions. As there is a dearth of evidence and theory on dynamic scoring design, I take the dynamic quality incentives in MA as fixed. Extending this analysis to incorporate market dynamics would be helpful for policy design. Also, I assume that quality is accurately measured and timely. Incorporating informational frictions to the designer's problem, and more importantly, data manipulation as in Ball (2019) would help extend these tools to markets where that has been an issue, such as nursing homes (Silver-greenberg and Gebeloff, 2021). Finally, I assume that the quality domain and dimensions are fixed. Allowing investments to exceed observed qualities or the designer to choose from new dimensions might also prove valuable.

References

Abaluck, J., Caceres Bravo, M., Hull, P., and Starc, A. (2021). Mortality Effects and Choice Across Private Health Insurance Plans. *The Quarterly Journal of Economics*, 136(3):1557–1610.

Aizawa, N. and Kim, Y. S. (2018). Advertising and risk selection in health insurance markets. *American Economic Review*, 108(3):828–867.

Albano, G. L. and Lizzeri, A. (2001). Strategic certification and provision of quality. *International Economic Review*, 42(1):267–283.

Allende, C., Gallego, F., and Neilson, C. (2019). Approximating the Equilibrium Effects of Informed School Choice. *Working Paper*, (1100623).

Angrist, J. D. and Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, 27(5):483–503.

Anscombe, F. J. and Aumann, R. J. (1963). A Definition of Subjective Probability. *The Annals of Mathematical Statistics*, 34(1):199–205.

Araya, S., Elberg, A., Noton, C., and Schwartz, D. (2018). Identifying Food Labeling Effects on Consumer Behavior. *SSRN Electronic Journal*.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*.

Atal, J., Cuesta, J. I., and Saethre, M. (2021). Quality Regulation and Competition: Evidence from Pharmaceutical Markets.

Baker, A., Larcker, D. F., and Wang, C. C. Y. (2021). How Much Should We Trust Staggered Difference-In-Differences Estimates? *SSRN Electronic Journal*, (March).

Ball, I. (2019). Scoring Strategic Agents. (January):1–63.

Barahona, N., Otero, C., Otero, S., and Kim, J. (2020). Equilibrium Effects of Food Labeling Policies.

Barrios, J. M. (2017). Occupational Licensing and Accountant Quality: Evidence from Linkedin. *SSRN Electronic Journal*.

Berry, S., Eizenberg, A., and Waldfogel, J. (2016). Optimal product variety in radio markets. *RAND Journal of Economics*, 47(3):463–497.

Berry, S. and Haile, P. (2020). Nonparametric Identification of Differentiated Products Demand Using Micro Data. *National Bureau of Economic Research*.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, 63(4):841.

Berry, S. T. (1994). Estimating Discrete-Choice Models of Product Differentiation. *The RAND Journal of Economics*, 25(2):242.

Berry, S. T. and Waldfogel, J. (2001). Do mergers increase product variety? Evidence from radio broadcasting. *Quarterly Journal of Economics*, 116(3):1009–1025.

Boleslavsky, R. and Kim, K. (2018). Bayesian Persuasion and Moral Hazard. *SSRN Electronic Journal*, (March).

Bollinger, B., Leslie, P., and Sorensen, A. (2011). Calorie posting in chain restaurants. *American Economic Journal: Economic Policy*, 3(1):91–128.

Brown, J., Duggan, M., Kuziemko, I., and Woolston, W. (2014). How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage Program. *American Economic Review*, 104(10):3335–3364.

Brown, Z. Y. (2018). An Empirical Model of Price Transparency and Markups in Health Care. *Working Paper*, (August).

Charbi, A. (2020). The fault in our stars! Quality Reporting, Bonus Payments and Welfare in Medicare Advantage*.

Chen, Y. (2018). User-Generated Physician Ratings-Evidence from Yelp.

Chernew, M., Gowrisankaran, G., and Scanlon, D. P. (2008). Learning and the value of information: Evidence from health plan report cards. *Journal of Econometrics*, 144(1):156–174.

Chou, S. Y., Deily, M. E., Li, S., and Lu, Y. (2014). Competition and the impact of online hospital report cards. *Journal of Health Economics*, 34(1):42–58.

Clay, K., Severnini, E., and Sun, X. (2021). Does LEED Certification Save Energy? Evidence from Federal Buildings.

CMS (2016). Quality Strategy. Technical report.

Crawford, G. S., Shcherbakov, O., and Shum, M. (2019). Quality overprovision in cable television markets †. *American Economic Review*, 109(3):956–995.

Crawford, G. S. and Shum, M. (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica*, 73(4):1137–1173.

Curto, V., Einav, L., Finkelstein, A., Levin, J., and Bhattacharya, J. (2019). Health care spending and utilization in public and private medicare. *American Economic Journal: Applied Economics*, 11(2):302–332.

Curto, V., Einav, L., Levin, J., and Bhattacharya, J. (2021a). Can health insurance competition work? Evidence from medicare advantage. *Journal of Political Economy*, 129(2):570–606.

Curto, V., Einav, L., Levin, J., and Bhattacharya, J. (2021b). Can health insurance competition work? Evidence from medicare advantage.

Dafny, L. and Dranove, D. (2008). Do report cards tell consumers anything they don't already know? The case of Medicare HMOs. *RAND Journal of Economics*, 39(3):790–821.

Darden, M. and McCarthy, I. M. (2015). The star treatment: Estimating the impact of star ratings on medicare advantage enrollments. *Journal of Human Resources*, 50(4):980–1008.

Dranove, D. and Dafny, L. (2008). Do Report Cards Tell Consumers Anything They Don ′ T Already. *RAND Journal of Economics*, 39(3):790–821.

Dranove, D. and Jin, G. Z. (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature*, 48(4):935–963.

Dranove, D. and Sfekas, A. (2008). Start spreading the news: A structural estimate of the effects of New York hospital report cards. *Journal of Health Economics*, 27(5):1201–1207.

Dworczak, P. and Martini, G. (2019). The simple economics of optimal persuasion. *Journal of Political Economy*, 127(5):1993–2048.

Elfenbein, D. W., Fisman, R., and McManus, B. (2015). Market structure, reputation, and the value of quality certification. *American Economic Journal: Microeconomics*, 7(4):83–108.

Fan, Y. (2013). Ownership consolidation and product characteristics: A study of the US daily newspaper market. *American Economic Review*, 103(5):1598–1628.

Fan, Y. and Yang, C. (2020). Competition, product proliferation, and welfare: A study of the US smartphone market. *American Economic Journal: Microeconomics*, 12(2):99–134.

Farronato, C., Fradkin, A., Larsen, B., and Brynjolfsson, E. (2020). Consumer Protection in an Online World: An Analysis of Occupational Licensing. *National Bureau of Economic Research Working Paper Series*, No. 26601.

Feng Lu, S. (2012). Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes. *Journal of Economics and Management Strategy*, 21(3):673–705.

Forbes, S. J., Lederman, M., and Tombe, T. (2015). Quality disclosure programs and internal organizational practices: Evidence from airline flight delays. *American Economic Journal: Microeconomics*, 7(2):1–26.

Frank, R. G. and McGuire, T. G. (2019). Market Concentration and Potential Competition in Medicare Advantage. *Issue brief (Commonwealth Fund)*, 2019(February):1–8.

Gandhi, A., Froeb, L., Tschantz, S., and Werden, G. J. (2008). Post-merger product repositioning. *Journal of Industrial Economics*, 56(1):49–67.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Goolsbee, A. and Petrin, A. (2004). The consumer gains from direct broadcast satellites and the competition with cable TV. *Econometrica*, 72(2):351–381.

Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, 103(7):2643–2682.

Harbaugh, R. and Rasmusen, E. (2018). Coarse grades: Informing the public by withholding information. *American Economic Journal: Microeconomics*, 10(1):210–235.

Holmstrom, B. and Milgrom, P. (1991). Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization*, 7(special_issue):24–52.

Hopenhayn, H. and Saeedi, M. (2019). Optimal Ratings and Market Outcomes.

Horowitz, J. L. and Markatou, M. (1996). Semiparametric Estimation of Regression Models for Panel Data. *Review of Economic Studies*, 63(1):145–168.

Houde, S. (2018a). Bunching with the Stars: How Firms Respond to Environmental Certification. *SSRN Electronic Journal*, (July).

Houde, S. (2018b). The Incidence of Coarse Certification: Evidence from the Energy Star Program. *SSRN Electronic Journal*.

Hui, X., Saeedi, M., Spagnolo, G., and Tadelis, S. (2018). Certification, Reputation, and Entry: An Empirical Analysis. *NBER Working Paper*, 24916.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831.

Jin, G. Z. and Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *Quarterly Journal of Economics*, 118(2):409–451.

Jin, Y. and Vasserman, S. (2019). Buying Data from Consumers.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kleiner, M. and Soltas, E. (2019). A Welfare Analysis of Occupational Licensing in U.S. States. *National Bureau of Economic Research Working Paper Series*, (1122374).

Krasnokutskaya, E. (2011). Identification and estimation of auction models with unobserved heterogeneity. *Review of Economic Studies*, 78(1):293–327.

Kuo, F. Y. and Nuyens, D. (2016). A practical guide to quasi-Monte Carlo methods. (November 2016):1—-50.

Larsen, B. (2014). Occupational Licensing and Quality: Distributional and Heterogeneous Effects in the Teaching Profession. *SSRN Electronic Journal*, (0645960):1–52.

Larsen, B., Ju, Z., Kapor, A., and Yu, C. (2020). THE EFFECT OF OCCUPATIONAL LICENSING STRINGENCY ON THE TEACHER QUALITY DISTRIBUTION. *National Bureau of Economic Research Working Paper Series*.

Lustig, J. (2010). Measuring welfare losses from adverse selection and imperfect competition in privatized medicare. *Manuscript. Boston University Department of . . .* , pages 1–49.

McGuire, T. G., Newhouse, J. P., and Sinaiko, A. D. (2011). An economic history of Medicare Part C. *Milbank Quarterly*, 89(2):289–332.

McManus, B. (2007). Nonlinear pricing in an oligopoly market: The case of specialty coffee. *RAND Journal of Economics*, 38(2):512–532.

Medicare Payment Advisory Commission (2013). The Medicare Advantage Program: Status Report. *Report to Congress: Medicare Payment Policy*, pages 287–306.

MedPAC (2020). The Medicare Advantage Program: Status Report. *Report to Congress: Medicare Payment Policy*, pages 287–306.

Miller, K. S., Petrin, A., Town, R., and Chernew, M. (2019). Optimal Managed Competition Subsidies. *National Bureau of Economic Research Working Paper Series*, No. 25616.

Mizala, A. and Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics*, 103(1):313–335.

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

Neal, D. and Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92(2):263–283.

Newhouse, J. P. and McGuire, T. G. (2014). How successful is medicare advantage? *Milbank Quarterly*, 92(2):351–394.

Nosal, K. (2011). Estimating Switching Costs for Medicare Advantage Plans.

Nosko, C. (2014). Competition and quality choice in the cpu market. *Manuscript, Harvard University*, (November):1–54.

Reid, R. O., Deb, P., Howell, B. L., and Shrank, W. H. (2013). Plan Star Ratings and Enrollment. *Journal of the American Medical Association*, 309(3):267–274.

Reyes, M., Garmendia, M. L., Olivares, S., Aqueveque, C., Zacarías, I., and Corvalán, C. (2019). Development of the Chilean front-of-package food warning label. *BMC Public Health*, 19(1):1–11.

Rodina, D. and Farragut, J. (2016). Inducing Effort Through Grades. *Working Paper*.

Ronnen, U. (1991). Minimum Quality Standards , Fixed Costs , and Competition Author (s): Uri Ronnen Published by : Wiley on behalf of RAND Corporation Stable URL : http://www.jstor.org/stable/2600984 Minimum quality standards , fixed costs , and competition. *The RAND Journal of Economics*, 22(4):490–504.

Schennach, S. M. (2016). Recent Advances in the Measurement Error Literature.

Schmalensee, R. (1979). Market structure, durability, and quality: a selective survey. *Economic Inquiry*, XVII.

Silver-greenberg, J. and Gebeloff, R. (2021). Maggots, rape and yet five stars: How u.s. ratings of nursing homes mislead the public. The New York Times https://www.nytimes.com/2021/03/13/business/nursing-homes-ratings-medicare-covid.html. Accessed: 06/26/2021.

Small, K. A. and Rosen, H. S. (1981). Applied Welfare Economics with Discrete Choice Models. *Econometrica*, 49(1):105.

So, J. (2019). Adverse Selection, Product Variety, and Welfare.

Spence, A. M. (1975). Monopoly , Quality , and Regulation. *The Bell Journal Of Economics*, 6(2):417–429.

Sweeting, A. (2009). The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria. *RAND Journal of Economics*, 40(4):710–742.

Town, R. and Liu, S. (2003). The Welfare Impact of Medicare HMOs. *The RAND Journal of Economics*, 34(4):719.

Train, K. (2015). Welfare calculations in discrete choice models when anticipated and experienced attributes differ: A guide with examples. *Journal of Choice Modelling*, 16:15–22.

Werner, R. M., Norton, E. C., Konetzka, R. T., and Polsky, D. (2012). Do consumers respond to publicly reported quality information? Evidence from nursing homes. *Journal of Health Economics*, 31(1):50–61.

Zapechelnyuk, A. (2020). Optimal Quality Certification. *American Economic Review: Insights*, 2(2):161–176.