Quality Disclosure and Regulation: Scoring Design in Medicare Advantage *

Benjamin Vatter†

April 30, 2024

Abstract

Policymakers and market intermediaries often use quality scores to alleviate asymmetric information about product quality. Scores affect the demand for quality and, in equilibrium, its supply. Equilibrium effects break the rule whereby more information is always better, and the optimal design of scores must account for them. In the context of Medicare Advantage, I find that consumers' information is limited, and quality is inefficiently low. A simple design alleviates these issues and increases total welfare by 3.7 monthly premiums. More than half of the gains stem from scores' effect on quality rather than information. Scores can outperform full-information outcomes by regulating inefficient oligopolistic quality provision, and a binary certification of quality attains 98% of this welfare. Scores are informative even when coarse; firms' incentives are to produce quality at the scoring threshold, which consumers know. The primary design challenge of scores is to dictate thresholds and thus regulate quality.

Keywords: disclosure, quality regulation, information design, equilibrium effects, welfare, competition
JEL Codes: L15, L11, I11, I18, D82, D83

---

*First version: October 2021. I 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, Samuel Goldberg, Eilidh Geddes, Piotr Dworczak, Hugo Hopenhayn, Philip Haile, and seminar participants at Northwestern University and several other institutions for their valuable comments and suggestions. This work benefited from generous funding from the Robert Eisner Graduate Fellowship. Any errors are my own.

†Massachusetts Institute of Technology, email: bvatter@mit.edu

I Introduction

Quality scores are ubiquitous. From car emissions to school performance, regulators and certifying agencies rely on scores for disclosure. Scores help consumers choose when information is scarce and, by doing so, also alter firms' incentives to invest in quality. While a growing theoretical literature provides valuable guidelines for designing disclosure policies, their welfare-optimal design depends on empirical fundamentals such as consumers' willingness to pay for quality, the degree of quality competition, and firms' ability to adjust to disclosure. The wrong design can exacerbate information frictions, distort firms' incentives, and even harm consumers.[1] Due to such concerns, much of the empirical literature on disclosure evaluates scores' ambiguous impact. In this paper, I bridge theory and empirics by studying optimal design in a real-world setting, estimating its primitives, examining the gains from alternate designs, and quantifying the relative importance of different design choices.

The effect of scores on the supply of quality breaks the rule that more information is always better for consumers (Blackwell, 1953). Coarser information can benefit consumers by regulating inefficiencies in quality provision caused, for example, by externalities in R&D, production subsidies, or limited competition. This paper focuses on the latter, specifically, on *Spencian* distortions (Spence, 1975) caused by firms' inability to capture surplus created by marginal quality increments from inframarginal consumers. When these distortions are not competed away, scores can coordinate demand to penalize inefficient firms and, simultaneously, reveal to consumers whether products are of efficient quality. Hence, in equilibrium, scores can lead to efficiency in quality and information.

I apply these ideas to study the Medicare Advantage (MA) Star Rating health insurance scores. This policy assigns plans a score between 1 and 5 stars, in half-star increments, according to their performance along five quality dimensions.[2] The MA setting provides a valuable laboratory for studying disclosure design: The rules mapping quality measurements to scores—i.e., the scoring design—vary annually, the regulator's quality measurement data are readily available for all plans, and there are no competing sources of quality scores for consumers. Moreover, firms are incentivized to compete on quality because their revenue is risk-adjusted, and premiums are highly regulated and subsidized. It is also an important setting in its own right: There are over 65 million Medicare beneficiaries, and quality impacts mortality (Abaluck *et al.*, 2021) and entails billions in public spending (CMS, 2016).

[1] For an example of harm in the nursing home industry see Silver-greenberg and Gebeloff (2021).

[2] These are preventive care, access to care, medical quality, chronic condition management, and patient experience. I detail the design in Section III.

I document three fundamental observations about scoring design in MA. In Section IV, I show that, first, consumers have increasing preferences for scores: New enrollees are 20% more likely to choose a 5-star than a 2-star plan, all else equal. Second, consumers' preferences correlate with changes to the mapping between quality and scores: The preference for 5 over 2-star plans depends on which qualities are awarded 5 instead of 2 stars. Third, firms respond to design changes by adjusting quality rapidly and proportionally to the scoring incentives. These observations and the variation that underlies them reveal consumers' preferences for scores and firms' quality production costs. Through the lens of a model, these can predict the impact of counterfactual scoring policies.

In Section V, I develop a model of quality investment, plan pricing, and enrollment, capturing four key frictions. First, consumers cannot distinguish between the qualities of equally rated plans. Second, unless the scoring design aggregates quality dimensions precisely as consumers' preferences, consumers cannot tell whether a higher-rated plan has a preferred aggregate quality over a lower-rated one. Third, firms have market power over price and quality, leading them to potentially inefficient investment and pricing decisions (Crawford *et al.*, 2019). Fourth, since consumers cannot ascertain by which combinations of qualities a plan obtained its score, firms ignore consumers' preferences when deciding how to allocate investments across quality dimensions.

The first two frictions present the designer with an opportunity to increase efficiency by improving the informativeness of scores. The other two frictions introduce a potentially opposite pressure to regulate investment moral hazard by coarsening scores. As firms' incentives are to attain scores at the lowest cost, investments target scoring thresholds. Thus, the number of scores controls the variety of qualities offered in the market (Kolotilin and Zapechelnyuk, 2019). Adding granularity to the design improves information and allows consumers of heterogeneous willingness to pay (WTP) for quality to match with diverse products, but also increases the potential for inefficient quality production.

In Section VI, I show that the primitives underlying the informational and moral hazard frictions are identified from variation in MA's design, enrollment, and quality. Consumers' WTP for scores is identified from the trade-off between premiums and scores in enrollment. Their preferences and beliefs about plan quality are identified from the correlation of WTP and changes to the scoring design. The same variation identifies firms' investment costs because it changes the relative gains of investing in different quality dimensions.

Model estimates reveal that quality is inefficiently provided and consumers' information is limited. A marginal improvement in the average contract's quality increases consumers' surplus between $17 and $84 million more than it costs to produce, depending on the quality

dimension. Quality is more efficiently provided in more competitive markets and when it is better represented in the scoring design. On the consumers' side, information frictions reduce their surplus by approximately four monthly premiums. Consumers' inability to discern if higher-scoring products have a preferred overall quality accounts for 94.5% of this loss since 22.7% of plans are *misclassified* from consumers' perspective.

I use the model estimates and a novel methodology to find an alternative, constrained optimal design for MA.[3] The new system is a simple discretization of plans' weighted average qualities into four scoring levels (five fewer than the Star Ratings) with three key features. First, medium-to-low qualities are pooled at the bottom score. Pooling decreases consumers' expectations of plan quality and induces a demand penalty for underprovision, which lessens the Spencian distortion. Second, more scoring levels are assigned to higher qualities, which balances product variety and efficiency and reduces within-score informational frictions. Third, the averaging weights are optimized to align with consumers' preferences, eliminating across-score frictions and multitasking moral hazards. This final feature is the most important; a binary certification with optimal weights attains 98% of the constrained optimum's welfare. Thus, the granularity of scores—their most visible and discussed design choice—is the least welfare-relevant once we account for equilibrium supply responses.

The alternative increases consumer surplus by $47.9 per beneficiary year and total welfare by $155.7. Design changes improve consumers' information, increase product quality, and increase prices. The mean squared error of consumers' beliefs about quality decreases by 75.5%, contributing $70.45 of the welfare gains. Average investment in quality nearly triples, contributing $90.14 of the welfare gains. A fraction of these gains are offset by a 3.8 p.p increase in insurance markups, as greater information and larger quality differentials across firms reveal and exacerbate vertical differentiation.

Quality regulation is the primary driver of the scores' welfare gains. Scores marshal demand and coordinate consumers to offset the distortionary forces skewing quality supply. This coordination can be achieved with disclosure policies that are simple and easy to understand, such as average quality certifications. These findings are also robust to various regulatory challenges, such as a limited understanding of the scoring policy by consumers, asymmetric information about firms' costs, or regulatory objectives that differ from total welfare. These results are fundamentally a consequence of the equilibrium effect of scores on quality, which overturns the dominance of full information. Welfare under the new scores is 17% larger than under full information.

[3]The constraint is to the space to which the Star Ratings belong. This is the class of all designs that deterministically assign a higher quality to weakly greater scores, using finitely many scoring levels.

Given the substantial gains provided by the alternative design, one might wonder why CMS's policies have systematically differed. One explanation is that CMS's objective includes factors beyond those considered here, such as the cost of future subsidized care. I consider the possibility of private regulatory preferences in Section VII.F, finding that CMS would have to value minor improvements in its preferred quality dimensions by exceedingly large values. A compelling alternative is discussed in Section VII.G, which considers the possibility that the regulator treats consumers as naive, with beliefs about quality that are independent and invariant to the scoring design. This naivety is testable and rejected by the data. However, if the regulator believes it and is extremely averse to misrepresenting consumers' preferences, the existing system outperforms the best simple scoring design. Given Medicare's delicate political and social role, these findings are, perhaps, reasonable.

Related Literature. This exercise in empirical scoring design bridges a gap between the theoretical literature on the subject and the empirical literature that measures disclosures' impact.[4] To my knowledge, few papers have explored this gap. Dai *et al.* (2018) study the optimal aggregation of subjective consumer restaurant reviews, while one of the counterfactual exercises in Barahona *et al.* (2022) explores optimal certification for ready-to-eat cereals. This paper extends these ideas to the broader agenda on information design with moral hazard (Boleslavsky and Kim, 2018) by examining optimal granularity, aggregation, and the trade-off between quality and informational regulation.

My results show that MA scores can act as effective quality regulation, which contributes to research on the supply effects of centralized mandatory disclosure (Jin and Leslie, 2003; Houde, 2018; Allende *et al.*, 2019; Barahona *et al.*, 2022) and the empirical study of quality regulation broadly (Angrist and Guryan, 2008; Kleiner and Soltas, 2019; Larsen *et al.*, 2020; Atal *et al.*, 2022). My examination of the regulation of imperfect competition among insurers expands on the literature on quality provision in healthcare markets (Cutler *et al.*, 2010; Cooper *et al.*, 2011; Gaynor *et al.*, 2013; Kolstad, 2013; Fleitas, 2020) and competition among insurers (Ho and Lee, 2017; Ho and Handel, 2021). In particular, I quantify the effects of moral hazard in quality provision among insurers competing for the demand of incompletely informed consumers. I prove that consumers' priors and quality preferences can be identified from design variation, which contributes to the study of choice under incomplete information

[4]On the theoretical side, these include Albano and Lizzeri (2001); Glazer and McGuire (2006); Harbaugh and Rasmusen (2018); Ball (2020); Hopenhayn and Saeedi (2019); and Zapechelnyuk (2020). The latter is particularly relevant, as it considers market power and moral hazard in scoring. On the empirical side, the literature includes Jin and Sorensen (2006a); Elfenbein *et al.* (2015); Araya *et al.* (2018); Alé-Chilet and Moshary (2022); Reynaert and Sallee (2021); and Charbi (2020), who measures the welfare value of the MA Star Ratings. See Dranove and Jin (2010) for a review of earlier work on quality disclosure and Kamenica (2019) for closely related work on theoretical information design.

(Abaluck and Gruber, 2011; Chernew *et al.*, 2008; Handel and Kolstad, 2015).

This paper connects research on the industrial organization of Medicare Advantage (Town and Liu, 2003; Lustig, 2009; Aizawa and Kim, 2018; Curto *et al.*, 2021; Nosal, 2011; So, 2019; Charbi, 2020; Miller *et al.*, 2022; Ryan, 2020; Decarolis *et al.*, 2020a) to the literature on insurance market design (Handel *et al.*, 2015; Marone and Sabety, 2022; Decarolis *et al.*, 2020b). I study the role of purely informational policies, whose implementation often focuses on statistical issues and maximizing informativeness. I provide evidence that their design must consider equilibrium supply effects and that doing so can drastically change the optimal solution. Closely related, Miller *et al.* (2022) study optimal subsidies and competition over coverage generosity in Medicare Advantage. This paper is complementary and extends the policy analysis to disclosure and competition over quality.

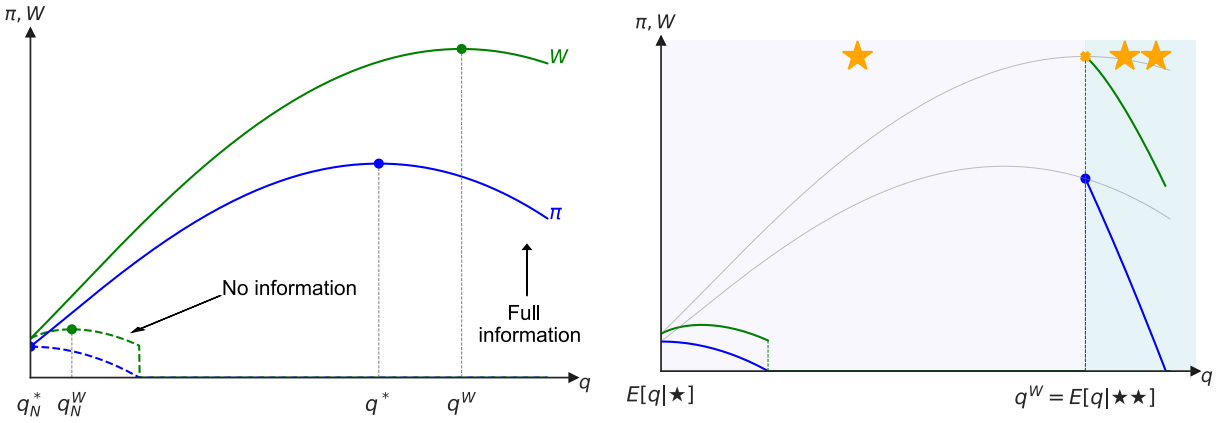
II Disclosure as Quality Regulation

Building on Spence (1975) and Zapechelnyuk (2020), I describe the economic intuition underlying scores' ability to regulate quality while informing consumers.[5] Consider a single-product monopolist selling an indivisible good. The monopolist chooses price and quality and pays a production cost that increases in its quality q. A regulator observes q and discloses a public score (or signal) $\psi(q)$. Consumers cannot observe q but know the scoring rule $\psi(\cdot)$ and the score. Using this information and knowing that quality is costly, consumers form rational expectations about q and make purchasing decisions. The regulator seeks to maximize welfare by committing to a policy before the monopolist's quality is chosen.

There are two informational extrema attainable by the scores. On the one hand, a constant score reveals no information to consumers, which renders demand inelastic to quality and thus eliminates any incentive for the monopolist to invest in it. On the other hand, a fully informative score allows the monopolist to exert market power over quality (Crawford *et al.*, 2019), leading to potentially inefficient investment (Spence, 1975). Intuitively, when evaluating a marginal increase in quality under full information, the monopolist considers its effect on the marginal consumer. The regulator, instead, accounts for the surplus created by the increase for marginal and inframarginal consumers. Hence, the monopolist's quality choice will likely be inefficient, even when consumers are fully informed.[6] Figure 1a illustrates these two extrema and the resulting inefficiencies.

[5] To focus on the primary mechanism used by the designer, I abstract away from unobserved investments, multidimensional quality, and other regulatory concerns found in the application.

[6] As noted by Spence (1975), this inefficient quality production can lead to over- or underprovision. Efficient output is also feasible under particular demand forms, such as linear.



(a) Informational extrema

(b) Regulated

Figure 1: Quality certification under monopolistic provision

Notes: These figures illustrate how certification changes a monopolist's investment incentives. Figure (a) presents profit and total welfare curves for the full and no-information scenarios. The latter vanishes when profits are negative, leading to a market exit. Figure (b) presents how profit and welfare change when consumers are only informed whether quality exceeds q^w. The monopolist's profit curve in (b) is disrupted as, within scoring intervals, consumers are not made aware of costly changes in quality, translating into no demand increases. Welfare is disrupted due to the fall in profits and because only consumers who buy the product regardless of quality improvements benefit from quality gains within intervals. Information is revealed at the threshold, restoring the curves to their original point. The welfare optimum in both figures is the same. Shaded areas illustrate the distinct scores.

The regulator can address these inefficiencies by using coarse scores. Figure 1b illustrates the outcome of a scoring rule that only certifies whether quality is at least efficient ($q \geq q^W$). This policy disrupts the firm's profit curve because, on both sides of the certification cutoff, demand is inelastic to quality but at different levels. To the left, consumers are guaranteed a ceiling on the quality of goods. Knowing that quality is costly to produce and that the monopolist lacks incentives to provide quality interior to the interval, they expect $q = 0$. To the right, consumers are guaranteed that quality is at least q^W and, by the same logic as before, expect $q = q^W$. Therefore, if the monopolist's full-information profits at $q = 0$ are lower than at $q = q^W$, it will invest efficiently when regulated. Consumers' expectations would then be accurate, thus eliminating market power over quality and informational distortions.

This illustration reveals that scores can improve on full-information outcomes by acting as quality-regulation policies. Their regulatory power stems from their ability to marshal demand to offset firms' market power. Scores can shift demand even if consumers have biased priors, face multiple products, or are subject to other sources of uncertainty. The solution often differs from a dichotomous certification since detailed scores accommodate product heterogeneity at the possible expense of decreasing firms' incentives to invest. To determine the optimal design for a given market, we must uncover firms' investment costs

and the social value of quality. In the following sections, I develop a methodology to recover these components and systematically translate them into optimal scoring designs.

III Institutional Details and Data

III.A Medicare Advantage and the Star Rating Program

Since 1965, retirees and disabled individuals in the US have had access to Medicare, a subsidized public health insurance system covering hospital, physician, and outpatient care. A series of reforms between 1982 and 2003 established an alternative to traditional Medicare (TM), known today as Medicare Advantage (MA). Under MA, the Centers For Medicare and Medicaid Services (CMS) contracts with private insurers to provide alternative coverage for Medicare beneficiaries in exchange for a prospective risk-adjusted capitated payment. Over the last decade, MA has become increasingly popular, covering 50.7% of the 65.9 million Medicare-eligible beneficiaries in 2024.[7]

MA markets are highly concentrated and regulated. In 2019, the average market (county) had 90% of its enrollment controlled by two firms. Nationally, four firms command 69% of all enrollment (Frank and McGuire, 2019). In most counties, insurers offer various plans that differ in coverage generosity (e.g., coinsurance) and access to clinical quality. CMS regulates the financial characteristics of plans, including minimum requirements on coverage generosity and limits on premiums relative to coverage (Curto *et al.*, 2021). CMS also subsidizes enrollees' premiums, resulting in zero premiums for nearly half of all MA plans.[8]

Differences in plan quality are less regulated and harder for consumers to ascertain. Quality varies across plans because of differences in the size and makeup of provider networks, disease management protocols, and processes for approving medical procedures, among other factors. Since insurers can offer the same network and services under different cost-sharing and premium combinations, CMS measures quality at the *contract* level. A contract is a group of plans from the same insurer that (according to CMS) share quality. The median contract has two plans, with 70% of its enrollment in one of them, and the median consumer observes only one of a contract's plans in her county's menu. Therefore, in many cases, the distinction between plan and contract is irrelevant. However, having price variation

[7]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 more information on risk adjustment and residual selection, see Brown *et al.* (2014) and So (2019).

[8]MA consumers pay a Part B premium regardless of their choice of TM or MA. For further details regarding the MA market regulation, see Appendix I.A

conditional on quality will play a role in the identification strategy. Throughout, I refer to products as "plans" and use the term "contract" only when relevant for clarity or exposition.

Information regarding plan quality is rarely available to insurance enrollees. To assist consumers, CMS created the Star Ratings scoring system, which displays a summary of each plan's quality next to the enrollment button in Medicare's unified shopping platform.[9] To compute these scores, CMS first collects information on over 60 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 1 to 5 to each plan measure, ascending in quality. Next, CMS chooses a weight for each category and computes a weighted average of all measure-level scores for each plan. Denoting w_k the weight of each category $k \in \mathcal{K}$ and $\mathcal{L}_k$ the measurements included in the category, the score of plan j is

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

Where $\text{Round}_{.5}(\cdot)$ rounds a number to its nearest half and q_{kj} is the quality of plan j in measure k. The adjustment factor, ω_j, captures minor bonuses due to past performance.[10]

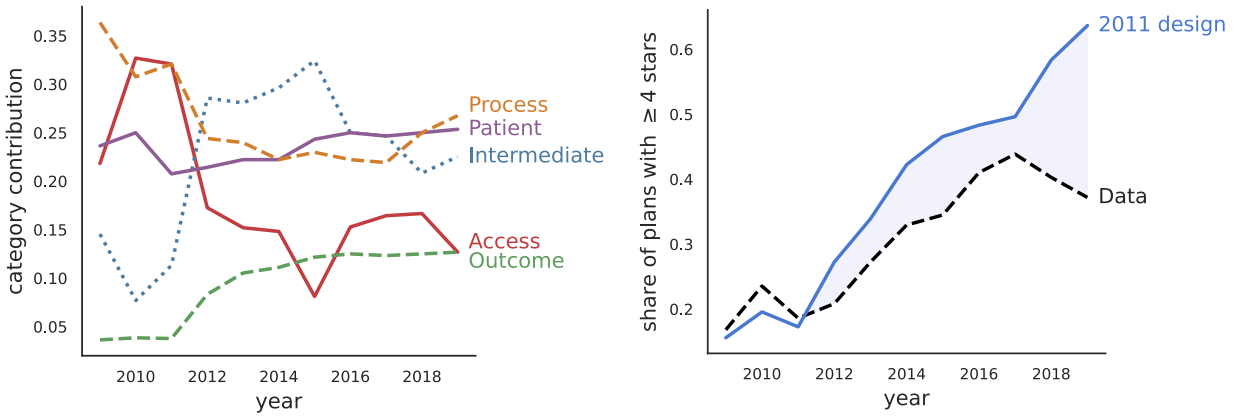
CMS frequently changed the weights and number of measures in each category, introducing substantial variation in the Star Rating design. In 2012, CMS moved from uniform weights to a design that gives each Outcome and Intermediate Outcome measure three times the weight of any Process measure and twice that of any Access or Patient Experience measure. The size of each category changed yearly as CMS experimented with measures. Given the high correlation across measures within a category, most of the analysis in this work is done at the category level.[11] I call the total weight assigned to a category its *design contribution*. These contributions have varied significantly, as shown in Figure 2a. As I detail in Appendix I.F, consumers likely observed this variation since the composition of categories was visible on the Medicare website and enrollment platform.

Design variation significantly impacted score assignment. Figure 2b shows that if CMS had kept its 2011 scoring design, 60% of plans in 2019 would have received 4 or more

[9] See Appendix I.B for a description of the online platform. For a description of earlier quality scores in MA, see Dafny and Dranove (2008).

[10] See the supplementary material for full construction details and a description of sources CMS uses to determine quality. Many of these measures are population and risk-adjusted, and very few come directly from insurers. See Appendix II.J for evidence against the influence of quality selection and manipulation.

[11] See the supplementary material for correlation within and across categories.



(a) Design variation

(b) Variation in score assignment

Figure 2: Scoring design variation and simulated assignment under constant design

Notes: Figure (a) shows the evolution of the scoring design category contributions. Each is the product of the category's 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. High correlation across measures within a category makes it natural to study design incentives at the category level and, thus, the design at the contribution rather than weight level. Figure (b) shows the change in the scoring assignment if CMS had kept its 2011 scoring design, keeping quality as measured in the data. The shaded area highlights the gap across the resulting assignments. Adjustment factors are preserved as measured.

stars, while the actual number was 40%. The difference is due primarily to a decrease in the importance of the Access category and an increase in the Outcome and Intermediate Outcome categories. Thus, in 2011, a high-scoring plan afforded consumers excellent access to physicians and a median-quality network of hospitals. 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 under a constant design.

Finally, CMS provides dynamic incentives. Starting in 2012, plan subsidies and scoring adjustment factors depend on past quality performance.[12] However, this paper aims to understand the short-run mechanisms, effects, and design of a purely informational quality disclosure policy. Thus, I incorporate dynamic features as they appear in the data and treat them as sources of heterogeneity. I exclude pecuniary incentives from the designer's toolkit to avoid confusing gains from information design with those from direct transfers.

[12]I ignore enrollment after the open enrollment period, which is allowed only for five-star plans. I also ignore contract consolidation, which few insurers exploited to manipulate their scores for a year.

III.B Data

This paper combines five data sources; the first is plan-market-level data from 2009 to 2019. Each year, CMS publishes every county's MA plans and their enrollment, subsidies, prices, rebates, premiums, plan benefits, and cost-sharing. The data provide the total number of Medicare-eligible beneficiaries in each county and information regarding the dual Medicare-Medicaid eligible population. I exclude dual eligibles and their plans from the analysis.[13]

The second source is the Medicare Current Beneficiary Survey (MCBS). This nationally representative rotating panel tracks around 15,000 Medicare beneficiaries for up to 4 years. I obtained data covering 2009 to 2015, which includes information on individual demographics, well-being, income, location, and enrollment.[14] The data includes linked medical claims and chronic condition information, which I use to compute each individual's risk score using CMS's risk adjustment software. In addition, I use the data to estimate each individual's predicted spending across all categories of care, including those not captured by CMS's risk scoring model.[15] I restrict the data to the continental US, leaving 46,833 beneficiary years. The panel also provides sampling weights to compare the survey's demographics with the national population. However, the data do not include all counties, limiting my analyses to about 22 million individuals or approximately one-third of the Medicare population.

The third source pertains to plans' quality and the scoring rules. CMS publishes the data used to compute the star ratings yearly, including quality measurements, assigned scores, and cutoffs. The data, however, do not explain changes to underlying measurement scales, weights, or variable definitions. To address this, I completed the data by reviewing a decade of CMS public communications aimed at insurers. I recovered year-to-year changes to the scoring design and replicated the public scoring assignment.

The fourth source corresponds to information about contract-level quality investment for 2015. The data comes from Medical Loss Ratio filings made by MA insurers, which recent regulation changes have modified to include a separate item for quality investment.[16] The fifth and final data source comes from the University of Wisconsin's County Health Rankings (Population Health Institute, 2024), which contains vital information about the availability

[13]Similar restrictions have been used by Aizawa and Kim (2018), Miller *et al.* (2022) and Curto *et al.* (2021). I present descriptive statistics in the Appendix Table 1.

[14]Excluding 2014, because it was never released to the public due to implementation difficulties.

[15]Another important difference between the risk scoring model and the predicted spending model is that the former uses substantially older data to assess both risk and spending. I provide details about these and all other data construction steps in the supplementary material.

[16]MLR regulation is not binding in MA and hence ignored during the analysis (Curto *et al.*, 2019).

of primary care physicians, population demographics, and other factors that might affect the cost of investing in quality in each county and the value of doing so for the local population.

IV Descriptive Evidence on Market Responses to Scoring

Scores' regulatory power stems from their effect on demand and, consequently, firms' investment incentives. However, whether the Star Ratings can influence demand and supply is an empirical question. For example, scores might be irrelevant if they summarize information consumers already know or fail to affect firms if their production technology is immutable. This section provides evidence that ratings affect both sides of the market.

IV.A Enrollment

Medicare beneficiaries' response to scores when making enrollment decisions has been thoroughly documented in previous work (Dafny and Dranove, 2008; Reid *et al.*, 2013; Darden and McCarthy, 2015).[17] This effect is easily observed in the individual enrollment panel by examining whether, all else equal, consumers prefer higher- to lower-scoring plans. Focusing on decisions made by new enrollees among MA plans, I regress a choice indicator on each plan's score, controlling for all observable (to consumers) plan attributes, including premiums, coverage, and additional benefits.[18]

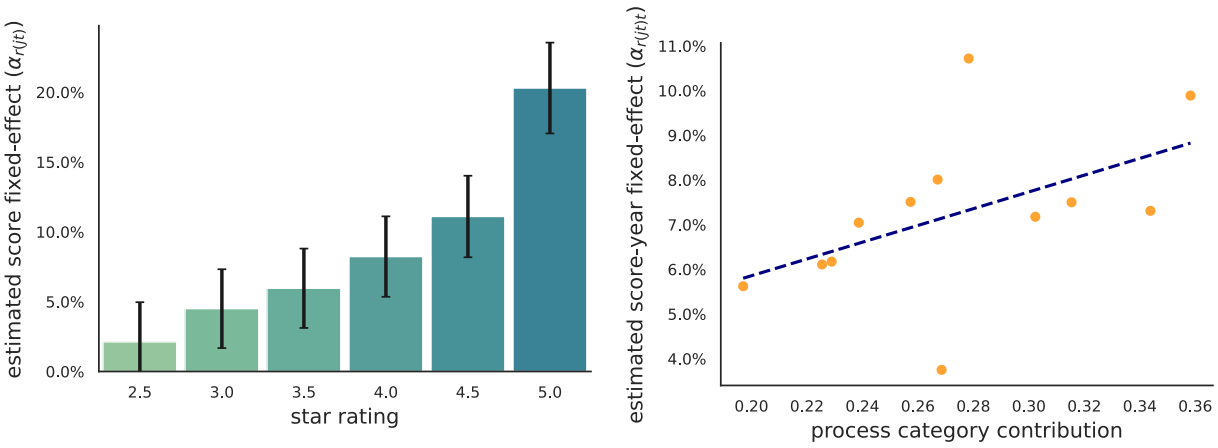
$$y_{ijt} = \alpha_{r(jt)} + \boldsymbol{x}_{jt}\boldsymbol{\lambda} + \mu_{m(i)} + \xi_t + \epsilon_{ijt} \tag{2}$$

Above, y_{ijt} indicates that consumer i chose plan j in year t, $\alpha_{r(jt)}$ is a fixed effect for plan j's score in year t, $\boldsymbol{x}_{jt}$ denotes plan characteristics, ξ_t a year fixed effect, and $\mu_{m(i)}$ a market fixed effect. Figure 3a displays the estimates of $\alpha_{r(jt)}$, the coefficients of interest.

All else equal, consumers prefer higher-scoring plans. A 5-star plan is approximately 20% more likely to be chosen by a new enrollee than an equivalent 2-star plan (the normalized category). Scores' effect on demand is monotonic, as expected if consumers understand that scores signal quality. However, whether consumers understand the informational content of scores is a different question. As I show in Section VII, even if consumers do not understand the nuances of the design, the regulator can produce effective scoring policies. Consumers have stated preferences for higher-rated products, and therefore, even if it stems from a naive

[17] The first two articles use aggregate enrollment data, while the third uses cross-sectional individual-level data. Here, I rely on individual-level panel data to select consumers potentially unaffected by inertia.

[18] As TM is not scored and inertia in plan choice is well documented for this market (Nosal, 2011), I focus on choices made by enrollees new to MA, conditional on choosing an MA plan.



(a) Conditional on design

(b) Across designs relative to Process category

Figure 3: Estimated effect of scores on plan choice probability, conditional on MA

Notes: Figure (a) displays the estimated $\alpha_{r(jt)}$ from equation (2). Error bars indicate 95% confidence intervals. The normalized category is two stars. The underlying coefficients and standard errors are presented in Appendix Table 2. Figure (b) shows the correlation between yearly estimates of consumers' preferences for scores and the contribution of the Process category. This is computed by estimating $\alpha_{r(jt)}$ from equation (2) separately by year and then residualizing each estimate against a fixed effect for the star-rating level. The figure shows a binned scatter plot of each residualized rating year against the category contribution. It illustrates the variation used for identification, as the approach infers preferences and beliefs from changes in choice likelihood within ratings as a function of scoring design.

understanding of the policy, the regulator can steer demand toward high-quality plans and change firms' investment incentives. Consumers' understanding, however, plays an important role in the identification strategy, as discussed in Section VI.[19]

To test whether consumers understand the policy variation, I conducted three exercises, detailed in Appendix I.F. The first and simplest exercise tests the null hypothesis that consumers are ignorant of any variation in the scoring design. To do so, it tests whether the likelihood of buying a high-rated plan changed following the redesign of 2012.[20] The results show a significant response, indicating that we can reject the null of total naivety to policy variation with a large degree of confidence. The second exercise asks whether the likelihood of buying a high-rated plan depends on the contribution of different quality categories in the scoring design. The results show that the contributions of all categories have a statistically significant effect on the choice likelihood. Finally, we can examine whether certain population

[19]Consumers' understanding affects the informational value of scores. However, the coordinating effect of scores can help them match with their preferred plans even if they face challenges in understanding the scoring design. This limits the losses from restricting the information channel, as discussed in Section VII.

[20]The 2012 design changes were particularly large and well-documented in the news.

groups respond more to ratings when additional weight is placed on quality categories that are relevant to them. The results show that consumers with chronic conditions are more likely to buy high-rated products than their healthy counterparts when the Intermediate Outcomes category contribution is higher. As this category summarizes important information about chronic condition management, this choice differential is consistent with consumers being aware of some of the changes in scoring design. Additionally, the results show that diabetic consumers are more likely to buy high-rated products when measures specific to diabetes management are more represented in the design.

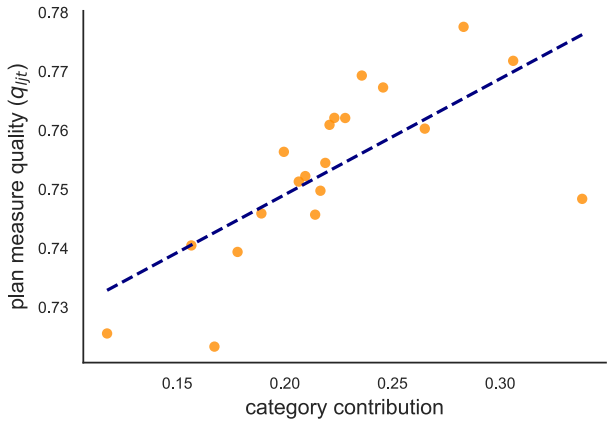
The results indicate that consumers value higher quality, as signaled by the scores. They are more likely to buy higher-scoring products to different extents depending on how the scores are designed. The demand model estimated in the main analysis relies on this joint variation between consumers' preferences for scores and the scoring design. Figure 3b illustrates one of the margins of variation relevant for identification, corresponding to the correlation between consumers' preferences and the Process category contribution. This correlation, along with that of other categories, determines the estimated ability of the regulator to influence choices through scoring design. Finally, Appendix I.G shows consumers are not wrong in valuing higher-scoring products as they are associated with better performance.

IV.B Quality

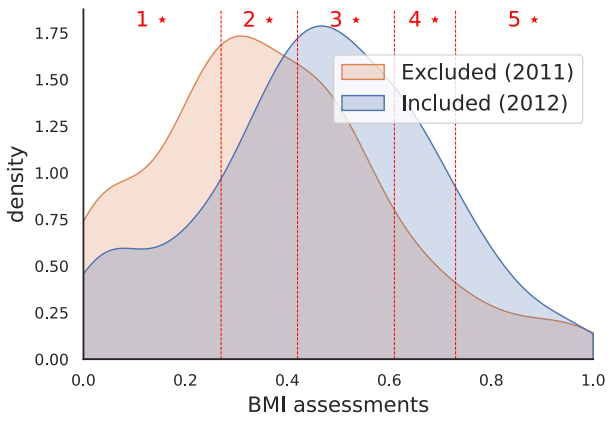
The first suggestive evidence that quality responds to scoring incentives is its correlation with category contributions (i.e., total category weight in the design) shown in Figure 4a. It illustrates how plan quality in any measure positively relates to its category's contribution. The figure isolates quality variation within plans, illustrating the extent to which a plan can vary its quality in response to scoring design.

To explore the causal link, I examine insurers' responses to the introduction of new quality measures to the design. This variation is a small subset of the factors changing category contributions but has three advantages. First, CMS evaluated the quality of these measures before their introduction. Second, these changes were announced to insurers without anticipation.[21] Finally, because the scoring rule converts quality measurements to measure-level scores, the change produced clear and heterogeneous incentives across firms. For example, Figures 4b and 4c show the distribution of two measures introduced in 2012 and 2018, respectively. In the first example, plans with a quality of 0.1 in 2011 faced the risk of getting 1 added to their list of measure-level scores if they failed to improve by 2012. As these scores

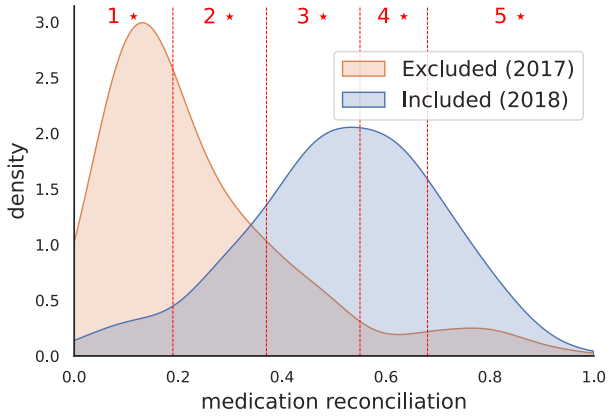
[21] Changes were announced a year before measurement, allowing insurers to respond in time.



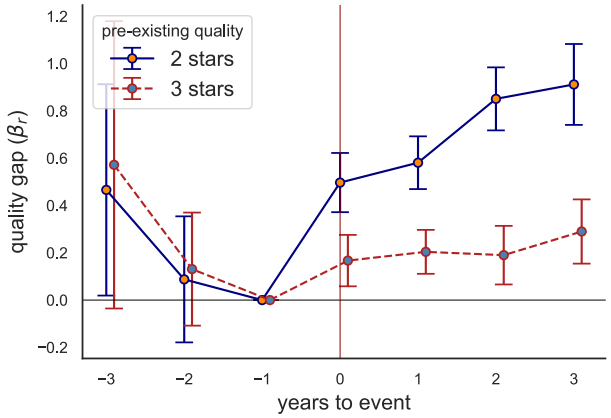
(a) To changes in category contribution



(b) To measure introduction, example 1



(c) To measure introduction, example 2



(d) To measure introduction, event study

Figure 4: Plan quality response to design variation

Notes: Figure (a) is a binned scatter plot of quality at the contract measure year level and its correlation with each measure's category contribution to that year's design. Observations have been residualized against a plan-measure fixed-effect to isolate variation within a plan over time. Figures (b) and (c) display the distribution of quality for two measures introduced to the design during the study period. Vertical lines mark the measure-level scoring bins in the introduction year. The horizontal axis marks the frequency with which a plan performs the quality process. Figure (d) shows estimates of equation (3), relative to the 4-star category and the year before introduction. Bars mark 95% confidence intervals.

are averaged over to form the star ratings, a failure to react would likely translate into a lower rating and, thus, a lower demand. In contrast, those with preexisting quality above 0.7 had no such incentive, as their measure-level score in 2012 would be five regardless. This logic applies to the second example and seven other such events.

I apply this logic to compare the evolution of quality across quality measures, plans, and time using a triple-difference regression. I assume preexisting heterogeneity in excluded quality measures was independent of the unanticipated change in design and that firms were

on similar trends across the thresholds. Therefore, plans of high preexisting quality follow the trend that low preexisting quality ones would have followed if not for the change in design. For plan j, quality measure l, and year t, I estimate the regression:

$$\underbrace{q_{ljt}}_{\text{normalized quality}} = \sum_{\tau=-3}^{3}\sum_{r=2}^{4} \underbrace{\beta_{r\tau}\mathbb{1}\{G_{lj}=r\}}_{\text{preexisting quality group}} + \underbrace{\gamma_{lj}+\mu_{lt}+\xi_{jt}}_{\text{pairwise fixed effects}} + \epsilon_{ljt} \tag{3}$$

Above, τ indexes time relative measure l' introduction and G_{lj} equals the measure-level score of each plan-measure using the design of the year of introduction ($\tau = 0$) applied to the quality of the preceding year ($\tau = -1$).[22] To compare quality metrics, I standardize them using their means and standard deviations across all years. To avoid conflating the effects of bounded quality domains, I drop plans in the first and last preexisting quality groups and normalize the coefficient of interest ($\beta_{r\tau}$) for the fourth group to zero.[23]

The analysis involves three differences. First is a comparison within plan-measure, controlled for by the fixed effect γ_{lj}. If only the post-indicator ($\mathbb{1}\{\tau \geq 0\}$) and this variable were included, then the coefficient on the indicator would reveal if, on average, quality increased following the design change. Second is a comparison across groups, captured by the groups (G_{lj}) and the measure-time fixed effect μ_{lt}. In this case, $\beta_{r\tau}$ would be positive for $\tau > 0$ and group r if their quality improved more after the change than for the comparison group ($r = 4$). Finally, the third difference compares across dimensions using plan-year fixed effects ξ_{jt}. The regression includes data on all quality measures, included or otherwise, such that this fixed effect accounts for the overall evolution of plan quality. Thus, the analysis compares quality changes in plan measures, accounting for quality trends in each dimension and plan. The coefficients of interest are identified from variation in quality within measures across time and its differential evolution across preexisting quality groups.

Figure 4d plots the $\beta_{r\tau}$ estimates, and Appendix I.H presents the underlying results and robustness to common methodological concerns. Before the design change, plans evolved similarly across the spectrum of preexisting quality. However, once incentives changed, plans of low preexisting quality improved substantially. Within a year, plans in the second group closed on average 29.6% of the gap between the 2-star and 5-star thresholds. Plans in the third closed 18.7% of their gap with five stars. In both cases, firms responded immediately; further improvements are minor and not statistically significant.

[22] For example, in Figure 4b, I classify a plan of measure quality 0.5 in the third group.

[23] The domain of most quality measures is bounded (e.g., the share of enrollees receiving a treatment). Therefore, low-quality plans can only improve, and high-quality ones can only worsen, and a failure to account for this would inflate the measurements of this analysis. See Appendix I.H for robustness.

Overall, the descriptive evidence reveals that consumers respond to scores by changing enrollment decisions and firms by adjusting quality. These adjustments are quick and vary depending on the stakes firms have in responding.[24] The following section presents a model that rationalizes these scoring effects and allows me to leverage MA's extensive variation. I identify consumers' preferences for plan attributes using variations in product characteristics and enrollment across markets and time. Changes in demand and subsidy rules perturb the marginal revenue of insurers, identifying their enrollment costs. Scoring design changes affect firms' incentives to invest in quality and reveal their costs. The same variation reveals consumers' valuation for scores, thus pinning their quality preferences and beliefs.

V Model

I model insurance provision and enrollment as the Perfect Bayesian equilibrium of repeated static interactions between consumers and insurers. Each year, the regulator discloses a national quality scoring rule. Insurers then simultaneously choose investments that stochastically determine plans' qualities. They then set plan prices, which subsidies and regulations convert to premiums and cost-sharing benefits. Finally, consumers observe premiums, benefits, and scores and enroll in TM or one of the MA plans available in their county. I present the game's stages in reverse order and discuss the model's central assumptions at the end.

V.A Demand

Building on Town and Liu (2003), 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 dental insurance), and a score of r_{jt}. Consumers maximize a Von Neumann-Morgenstern expected utility, evaluating subjective beliefs over quality. The expected indirect utility of consumer i from plan j is

$$u_{ijmt} = \underbrace{\alpha_i p^{\text{total}}_{jmt}}_{\text{premium}} + \underbrace{\beta_i b_{jmt}}_{\text{benefits}} + \underbrace{\mathcal{E}[v(\boldsymbol{q})|r_{jt},\psi_t]}_{\text{quality}} + \underbrace{\boldsymbol{\lambda}^{a\prime}\boldsymbol{a}_{jmt}}_{\substack{\text{plan}\\ \text{attributes}}} + \underbrace{\boldsymbol{\lambda}^{l\prime}\boldsymbol{l}_{ijt}}_{\substack{\text{demographic}\\ \text{preferences}}} + \underbrace{\xi_{jmt}}_{\substack{\text{unobserved}\\ \text{preference}}} + \underbrace{\varepsilon_{ijmt}}_{\sim T1EV} \tag{4}$$

Consumers have heterogeneous preferences for premiums and benefits (α_i, β_i). Following Curto *et al.* (2021), both variables are dollar-valued, with the latter being the expected dollars saved from insurance, according to CMS. Benefits summarize all cost-sharing attributes of

[24] CMS did not select measures randomly. Therefore, the results do not speak to the effect of scoring a generic quality dimension. The variation needed to measure these effects exists in the data, as scoring rules change yearly, yet disentangling them from the overall data variation requires a more structured approach.

the plan, such as copayments and coinsurance, and are shown on the enrollment platform. Total premiums include part C (MA) and D (prescription-drug) premiums associated with the plan.[25] Consumers value the vector of quality $\boldsymbol{q}$ at $v(\boldsymbol{q})$, forming subjective expectations about this vector given the scoring policy (ψ_t) and the plan's score (r_{jt}).[26] Consumers also value the plan's bundled services $(\boldsymbol{\lambda}^a)$ and have systematic preferences for certain insurers and MA overall based on their demographic group $(\boldsymbol{\lambda}^l)$. Finally, consumers have unobserved preferences for plans (ξ_{jmt}) and independent type-1 extreme value idiosyncratic preferences (ε_{ijmt}), as in Aizawa and Kim (2018) and Miller *et al.* (2022).

Consumers can also opt for TM coverage. Since most MA enrollees choose plans including prescription drug coverage, I assume they would also bundle TM with a Part D prescription drug plan. I denote by b_0 TM's standard insurance benefits and p^D_{0mt} the price of the market's most popular Part D plan.[27] Consumers' heterogeneous preferences for TM are captured by their demographic relative preferences for MA. The outside option's indirect utility is thus $u_{i0mt} = \alpha_i p^D_{0mt} + \beta_i b_0 + \varepsilon_{i0mt}$.

Given this model, the likelihood with which consumer i chooses product j in market m in year t is given by $s_{ijmt} = \frac{\exp(\delta_{ijmt})}{\exp(\delta_{i0mt})+\sum_{j'\in\mathcal{J}_{mt}}\exp(\delta_{ij'mt})}$ where $\delta_{ijmt} = u_{ijmt} - \epsilon_{ijmt}$, is the expected indirect utility of each option. Therefore, 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, $D_{jmt} = \sum_{i\in\mathcal{I}_{mt}} s_{ijmt}$. Each consumer is assigned an individual risk score γ_{it} for risk adjustment purposes. I denote the risk-adjusted demand of plan j as $\tilde{D}_{jmt} = \sum_{i\in\mathcal{I}_{mt}} \gamma_{it} s_{ijmt}$

V.B Supply

V.B.1 Pricing: Each year t, at the third stage of the game, insurance firm f observes the vectors of realized qualities $\boldsymbol{q}_t$ and scores $\psi_t(\boldsymbol{q}_t) = \boldsymbol{r}_t$. Given this information, the firm

[25] Premiums include all rate reductions. Consumers also pay a part B premium regardless of their choice, which cancels out. Benefits are shown to consumers as expected payments, while CMS evaluates these as insurer payments for regulatory purposes. I use the latter, so benefits increase with generosity.

[26] For micro-foundations of indirect utility with additive quality preferences, see Appendix II.B. The appendix also discusses how elements such as network quality and unmeasured quality dimensions are accounted for in the model.

[27] The only relevant characteristic of the outside option's part D plan is its price. Therefore, whether using the standard defined plan or the most popular plan is largely equivalent.

chooses prices to maximize its total profits given by[28]

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

The plan's marginal revenue per risk-adjusted consumer is the sum of its price (p_{jmt}) and additional revenue from prescription coverage and subsidies ($R(\cdot)$). The latter depends on the plan's price and attributes ($\boldsymbol{z}_{jt}$), including its counties of service and share of benefits financed with subsidies. I present the formula for this function and how prices map to premiums and benefits in Appendix II.A. Costs, $C(\cdot)$, covers a unit risk-score enrollee's standard Medicare benefits, prescription drugs, non-Medicare benefits (e.g., dental insurance), and management. This function varies according to the plan's quality ($\boldsymbol{q}_{jt}$), additional attributes as included in the demand ($\boldsymbol{a}_{jmt}$), and a set of unknown parameters to estimate ($\boldsymbol{\theta}^c$).

Premium and benefit regulations introduce a kink in the demand and revenue of a firm as a function of prices. If the firm sets prices above the kink (known as the benchmark), then a dollar price increase produces an equivalent increase in revenue and premiums, and cost-sharing is unaffected. Below the kink, a dollar increase in prices produces less than a dollar increase in revenue and premiums and a mandatory decrease in the plan's benefits.

V.B.2 Investment: In the game's second stage, each firm observes the regulator's scoring rule ψ_t and chooses an investment level x_{ckt} for each of its contracts $c \in \mathcal{C}_{ft}$ and category of quality k.[29] For example, an insurer can invest in forming networks with better providers to improve its Outcome quality or expand its network to improve Access quality. Firms' choices maximize their expected insurance profits net of the quality investment costs:

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

To form an expectation of its profits, firm f evaluates two dimensions of uncertainty. First, realized quality might differ from its intended target, captured by the conditional distribution $F(\boldsymbol{q}_f | \boldsymbol{x}_{ft})$. Second, firms are uncertain about their rivals' investment costs and, therefore, their choices at this stage. Since rival investments affect the firm's profits only insofar as they shift quality, firms take expectations over these realizations ($\boldsymbol{q}_{-f}$). I assume firms hold rational expectations over the distribution of rival qualities formed by observing market characteristics at investment time. These include their rivals' identity, consumers' demographic

[28] In MA, this price is called a "bid." I avoid this terminology to prevent confusion with auctions.

[29] Each contract is associated with a set of plans $\mathcal{J}_{ct}$ such that $\mathcal{J}_{ft} = \bigcup_{c \in \mathcal{C}_{ft}} \mathcal{J}_{ct}$.

characteristics, and their previous enrollment choices. The assumption is motivated by the secrecy of insurers' contractual arrangements and the lack of investment data.[30]

To understand insurers' investment problem and its interaction with the scoring policy, it is instructive to ignore investment risk. In this case, the problem consists of selecting an optimal rating for each plan and finding the cost-minimizing combination of qualities that attain this rating. Therefore, conditional on the target, the combination of qualities a contract has is independent of consumers' preferences: Consumers do not observe this combination, and thus, insurers ignore their preferences. Insurers only consider consumers' aggregate WTP for quality when choosing a target rating. As investment risk is independent of consumers' preferences, reintroducing it to the analysis does not change this intuition.

V.C Regulator

The regulator seeks to maximize a weighted sum of expected consumer surplus and insurer profit, net of governmental spending, by choosing a scoring policy ψ from within a class Ψ:

$$TW(\psi, \rho^F, \rho^G) = \int [\underbrace{CS(\psi, \boldsymbol{q})}_{\substack{\text{Consumer} \\ \text{surplus}}} + \rho^F \underbrace{\sum_f V_f(\psi, \boldsymbol{q}) - I_f(\boldsymbol{x}_f^*(\psi))}_{\text{Insurer profit}} - \underbrace{\rho^G G(\psi, \boldsymbol{q})}_{\substack{\text{Government} \\ \text{spending}}}] dF(\boldsymbol{q}|\boldsymbol{x}^*(\psi)) \quad (7)$$

Where $\boldsymbol{x}^*(\psi)$ denotes the equilibrium investment induced by the scoring policy, $G(\cdot)$ the governmental subsidy spending on TM services and MA enrollment, and $CS(\cdot)$ consumers' surplus from enrollment.[31] The regulator evaluates expected consumer surplus using the true rather than consumers' subjective distribution of quality. Thus, following Train (2015), consumers' surplus can be expressed as the sum of the ex-ante expected surplus (Small and Rosen, 1981) and an ex-post correction for the realization of quality:[32]

$$CS(\psi, \boldsymbol{q}) = \sum_{i \in \mathcal{I}} \frac{1}{|\alpha_i|} (\underbrace{\ln(\sum_{j \in \mathcal{J}_{mt} \cup \{0\}} \exp(\delta_{ijmt}))}_{\text{ex-ante surplus}} - \underbrace{\sum_{j \in \mathcal{J}_{mt}} s_{ijmt}(v(\boldsymbol{q}_{jt}) - \mathcal{E}[v(\boldsymbol{q})|\psi(q_{jt}), \psi])}_{\text{ex-post correction}}) \quad (8)$$

Where δ_{ijmt} depends implicitly on the realization of quality and scoring policy. As the regulator evaluates (8) taking expectations over realized qualities, maximizing consumers'

[30] I present evidence of imperfect quality control in Appendix I.J. The assumption about firms' beliefs is similar to that of Sweeting (2009). This article's limited investment data only recently became available.

[31] The governmental cost omits the part D subsidies, which would apply regardless of segment choice as consumers in the model get part D coverage, whether in TM or MA.

[32] This surplus standard is common in the literature on choice under uncertainty. Similiar approaches have been used by Jin and Sorensen (2006b), Allcott (2011), and more recently Reimers and Waldfogel (2021).

surplus consists of maximizing their ex-ante utility from enrollment net of the correlation between choices and expectational errors.

Scores can also help consumers compare across options, reducing the complexity involved with enrollment. As the data contains no meaningful variation in choice or scoring complexity, the main analysis addresses this gap by restricting counterfactual scoring policies to ones at most as complex as the status quo. Thus, when evaluating the effects of alternative systems, the net change in complexity should be close to zero. This missing component, however, might lead to substantially overstating the welfare value of complex informational environments like those induced by full information, as discussed further in Section VII.

The regulator is also concerned with the equilibrium effect of its scoring policy on firms' investment decisions and prices. The regulator faces two dimensions of moral hazard. First, when firms decide on target ratings for each contract, their objectives disagree with the regulators' on the value of providing quality to inframarginal consumers. This is the essence of the Spencian distortion, creating a principal-agent moral hazard problem (Holmstrom, 1982). Conditional on the rating, firms and the regulator also disagree on the combination of qualities to invest in. As noted above, insurers consider only their costs, while the regulator also accounts for consumers' preferences. This misalignment produces a multitasking moral hazard problem that further separates firm choices from socially optimal ones (Holmstrom and Milgrom, 1991). Overall, firms' endogenous investment response to the scores presents the regulator with a trade-off between providing consumers with information and forgoing the regulatory power of coarse scores as illustrated in Section II.

I do not impose any optimality on the regulator's policy decisions when estimating the model. CMS has been experimenting and responding to changes in Medicare policy. Thus, their scoring policy decisions likely reflect a combination of the above welfare objective and some implicit value from experimentation and satisfying stakeholders.

V.D Discussion

The model makes two simplifications that might affect the scoring design analysis. First, consumers have homogeneous preferences over quality, making it a vertical attribute (Mussa and Rosen, 1978). This reduces the computational cost of solving the scoring design problem, which is a stochastic optimization over a non-smooth functional space.[33] Appendix II.C

[33]The solution method's complexity is proportional to the product of the dimensions of quality, rival firms, quality shocks, and heterogeneity in consumer quality preferences. Thus, adding moderate heterogeneity can increase the time required to solve this problem from months to years. However, the method can solve the scoring design problem with heterogeneous quality preferences with fewer firms or quality dimensions.

examines this heterogeneity in the data, showing little evidence of meaningful differences across observable groups. Additionally, Appendix II.D shows that the modeled heterogeneity in WTP is sufficient to generate over- and underprovision of quality and thus capture fundamental market frictions. In Appendix III.E, I discuss two robustness exercises that speak to the role of preference heterogeneity in scoring design.

The second key simplification is that the game is static. Consumers do not learn from past experiences; firms do not carry over investments from previous years. Quality in MA, however, is primarily the outcome of contractual arrangements that change often and rapidly. The variation I document in Sections III and IV supports this claim. Moreover, the largest insurers in MA entered decades ago and have likely already invested in major components such as developing relationships with providers or software to track their populations' health. Therefore, dynamic investment incentives are likely to be second-order in this market. For consumers, the argument in favor of the assumption is similar.[34] To learn to predict future qualities, consumers would have to infer insurers' investment costs. As subsidies and scores mask the revenue and quality of contracts, this task would be challenging even for sophisticated consumers. Compounding with significant quality variation, this complexity makes it improbable that information acquired in a given year will be valuable in the next.

VI Identification and Estimation

VI.A Demand

I estimate the demand model using the two-step approach of Goolsbee and Petrin (2004). The first step uses individual-level enrollment decisions to recover preference heterogeneity and aggregate market shares to estimate mean population preferences. Splitting the premium and benefit parameters in equation (4) into their mean (α, β) and variation $(\tilde{\alpha}_i, \tilde{\beta}_i)$, the method aggregates mean preferences with all common components of a plan's utility—including quality—in a single scalar, δ_{jmt}. This transformation has three unknown components: preference heterogeneity $(\tilde{\alpha}_i, \tilde{\beta}_i)$, demographic preferences $(\boldsymbol{\lambda}^l)$, and plan-market-year fixed effects $(\boldsymbol{\delta})$. Collecting these in a vector $\boldsymbol{\vartheta}$, the first stage solves

$$\max_{\boldsymbol{\vartheta}} \underbrace{\sum_t \sum_i w_{it} \sum_{j \in \mathcal{J}_{m(i)t}} y_{ijmt} \ln(s_{ijmt}(\boldsymbol{\vartheta}))}_{\text{weighted log-likelihood}} \quad \text{s.t} \quad \underbrace{s^*_{jmt} = \sum_i w_{it} s_{ijmt}(\boldsymbol{\vartheta})}_{\text{share matching}} \quad \forall j, m, t \tag{9}$$

[34] Inertia hampers the separate identification of learning from systematic preferences. The rotating-panel structure of the MCBS further complicates this, as it follows consumers for only a few years.

Where y_{ijmt} is a choice indicator, $s_{ijmt}(\boldsymbol{\vartheta})$ is the model-implied individual choice probability, and s^*_{jmt} is the observed market share. Thus, the first step is a constrained weighted maximum likelihood problem, where w_{it} are nationally representative MCBS sampling weights. The constraint matches predicted and observed market shares, which I solve using the Berry (1994) inversion and the Berry *et al.* (1995) fixed-point contraction.

The second step is a two-stage least-squares regression of the estimated mean preferences on their components. I decompose consumers' unobserved preference (ξ_{jmt}) into systematic taste for MA in each market ($\mathbb{d}_{mt}$), systematic preferences for the contract ($\bar{\eta}_{c(j)}$), and all residual unobserved preference ($\tilde{\xi}_{jmt}$).

$$\hat{\delta}_{jmt} = \underbrace{\alpha p^{\text{total}}_{jmt}}_{\text{premium}} + \underbrace{\beta b_{jmt}}_{\text{benefits}} + \underbrace{\boldsymbol{\lambda}^{a\prime}\boldsymbol{a}_{jmt}}_{\text{plan attributes}} + \underbrace{\mathcal{E}[v(\boldsymbol{q})|r_{jt},\psi_t]}_{\text{quality}} + \underbrace{\bar{\eta}_{c(j)}}_{\text{contract FE}} + \underbrace{\mathbb{d}_{mt}}_{\text{market-year FE}} + \tilde{\xi}_{jmt} \quad (10)$$

Firms' knowledge of $\tilde{\xi}_{jmt}$ renders premiums, benefits, and scores endogenous in this regression. To address the endogeneity of premiums and benefits, I develop two instruments based on regulatory features of insurers' additional revenue ($R(\cdot)$). First, I use an average of TM's insurance cost in the plan's other markets. The regulation links each plan's subsidies with the public option's cost in every county where it participates, making the leave-one-out average a strong predictor of subsidies unaffected by local demand. Second, to distinguish between the effect of endogenous prices on premiums and benefits, I use variation across plans in the added revenue from pricing below the regulatory benchmark. Both instruments vary across plans and years due to county choices, regulations, and TM's cost variations.[35]

Consumers' unobserved preferences also influence firms' investments and, thus, scores. We can view consumer's preferences for quality and systematic preferences for contracts as a single endogenous contract-year fixed-effect $\eta_{c(j)t} = \mathcal{E}[v(\boldsymbol{q})|r_{jt},\psi_{jt}] + \bar{\eta}_{c(j)}$. I address this endogeneity by relying on instruments that interact the investment multitasking moral hazard problem with the scoring design variation. Formally, the instruments are the set $\{\frac{\omega_{kt}}{\omega_{k't}}\frac{q_{ckt}}{q_{ck't}}\}_{k,k'\in\mathcal{K}}$, where ω_{kt} is the contribution of category k in year t. The first ratio, $\omega_{kt}/\omega_{kt'}$, captures changes in the design that might benefit different firms. For example, if this ratio grows, firms with a cost advantage in providing k over k' should find it cheaper to obtain higher scores. The second ratio captures a contract's cost advantage. As is the essence of the multitasking moral hazard problem, firms' relative investment across dimensions is independent of consumers' preferences and governed primarily by their cost structure. Therefore,

[35]The exclusion restriction would fail if, for example, plans changed counties due to the correlation between TM cost and plan preference. As 92% of non-terminated plans remain in a county the following year, this seems unlikely.

the interaction between the ratios captures variations in the design that enhance or hamper different contracts' ability to obtain scores. The set of instruments excludes permutations of quality dimensions (i.e., if $q_k/q_{k'}$ is included, then $q_{k'}/q_k$ is not) and an arbitrary normalized pair, to not pin-down the aggregate quality level.[36] These instruments should satisfy the exclusion restriction as long as the regulators' design variation is exogenous to changes in consumers' unobserved preferences for specific plans.

Appendix II.E presents additional details about the instruments, evidence on the underlying source of the endogeneity, the instruments' first stage in the above regressions, and evidence that suggests that these instruments might satisfy the exclusion restriction.

VI.A.1 Quality beliefs and preferences: In estimation, consumers' preferences for scores are star-year fixed effects absorbed within $\eta_{c(j)t}$. Their separate identification from variation in plans' scores follows standard identification arguments (Berry and Haile, 2020). Intuitively, consumers reveal these preferences when trading off premium increases for rating changes. The challenge is that these valuations do not reveal consumers' preferences for quality separately from their beliefs. For example, consumers might be willing to pay a substantial amount for plans to have 4 instead of 3 stars, all else equal. This preference can be based on a belief that 4-star plans are of starkly superior quality or because consumers substantially value even slight differences in quality. Disentangling beliefs from preferences requires an assumption on how consumers form beliefs, given the scores they observe:

Assumption 1 (Consumer beliefs)**.** *One of the two hold: (1)* ***Informed choice****: Consumers know $\psi_t(\cdot)$ and use scores and Bayes' rule to update a continuous prior density $f : \mathcal{Q} \to \mathbb{R}_+$, with compact and connected support; (2)* ***Ignorance****: Consumer's posterior beliefs $\mathcal{E}[\boldsymbol{q}|r, \psi]$ are exogenous, independent of ψ, and bounded in $\mathcal{Q}$.*

Informed choice is the most common assumption in the literature. In theoretical work, consumers (receivers) often know precisely the rules by which the regulator (sender) transforms the distribution of quality (state) (Kamenica and Gentzkow, 2011). In the empirical literature, consumers either know the true structure or a parametric and unbiased approximation of it (Crawford and Shum, 2005; Dranove and Sfekas, 2008; Barahona *et al.*, 2022).[37] Crucially, in both cases, the econometrician knows how consumers interpret scores and can rely on their variation. In addition, consumers' knowledge of the scoring rule allows the regulator to shape their beliefs, which gives additional power to the scoring policy; the ignorance

[36] Even if all combinations were included, it would not fully predict a contract's rating as cutoffs vary.

[37] Alternatively, some allow for parametric bias based on additional data, such as external surveys.

assumption generates the other extreme of no informational power.[38]

These assumptions gain power once combined with appropriate variation in scores. In Appendix II.F, I show that since MA scores are a weighted average of quality partitions, they are well approximated within the class of *monotone partitional scores* (Dworczak and Martini, 2019). This class includes all scores that partition quality space into numbered partitions, assigning weakly greater labels to strictly greater quality.[39] Therefore, to score a plan, one only needs to assess in which partition its quality fell. This class of scores is exceedingly common and includes all deterministic certifications of quality (e.g., front-of-package nutrition labels), letter grades (e.g., restaurant hygiene scores), and many others.

Assumption 2. *(Design variation) ψ_t is drawn from a distribution with a strictly positive density over partitional scores with linear boundaries and $N \geq 3$ partitions, with N fixed.*

Assumption 2 states that scores will continue to vary within a set that includes but is not limited to, the type of designs observed in the data. It does not require that the number of partitions grow with the sample or entail complex aggregation rules (i.e., boundaries). The key identification result, proven in Appendix II.G, follows.

Proposition 1. *(Quality beliefs and preference identification) Let assumption 2 hold and quality preferences be linear, i.e., $v(q) = \boldsymbol{\gamma}'\boldsymbol{q}$. If assumption 1.a holds then $(\boldsymbol{\gamma}, f(\cdot))$ are identified. If assumption 1.b holds, then there is a nontrivial identified lower bound for $\boldsymbol{\gamma}$.*

This identification result depends on the setting only insofar as common consumer preferences for score-years can be identified, and the scoring design varies within a typical class.[40] Intuitively, consumers' willingness to pay for score increments implies bounds on their preferences and beliefs. For example, suppose quality is scalar, the prior is uniform, and $\gamma = 1$. If nine scores uniformly divide $[0, 1]$, consumers would be willing to pay $8/9$ more for a top-rated product than a bottom-rated one. Simple algebra shows that by observing differences in willingness to pay and knowing the scoring structure, we can bound γ within $(8/9, 8/7)$.

[38] Alternatively, consumers could hold rational expectations, implying that they know the scoring rule, firms' costs, and investment risk well enough to predict quality changes. Rational expectations would allow the regulator to control quality without informational losses, rendering informational policies stronger than those considered here.

[39] Dworczak and Martini (2019) consider a larger class of monotone partitional signals that fully reveal quality in some partitions. The order used also varies across applications.

[40] The result does not depend on the logit structure. Moreover, the restriction to linear partitions is immaterial. The full support assumption on scoring design is valuable for identifying the corners of prior beliefs. Still, the identification argument is not at the limit: Variation in design provides meaningful identifying restrictions even if all partitions have positive measures.

Scoring variation produces new intervals for γ, intersecting and shrinking the identified set down to a point. This process also bounds posterior beliefs and, thus, priors.

The result shows that informed choice is a powerful assumption. It imposes a strong structure on consumers' understanding and, in return, delivers identification. In MA, this assumption requires primarily that consumers know the relative contribution of categories to the scores, which is supported by the evidence of Section IV. Therefore, I rely on the informed choice assumption for the main analysis, leaving the ignorance assumption for robustness.

For results that rely on informed choice, I estimate preferences (now captured by a vector $\boldsymbol{\gamma}$) and prior beliefs ($f(\cdot)$) using a nonparametric minimum distance estimator. To remove any systematic preference for specific contracts, I only leverage time-series variation in consumers' preferences for contract years, $\eta_{c(j)t}$. The resulting estimator is

$$\min_{\boldsymbol{\gamma},\boldsymbol{\zeta}} \sum_{c(j)} \sum_{t} \sum_{\tau>t} \left(\Delta_t^\tau(\eta_{c(j)t} - \boldsymbol{\gamma}'\mathcal{E}[\boldsymbol{q}|r_{c(j)t}, \psi_t; \boldsymbol{\zeta}])\right)^2 \tag{11}$$

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

VI.A.2 Estimates: Panel A of Table 1 presents the estimated consumer premium and benefit preferences. A dollar in benefits is roughly equivalent to a \$2.25 reduction in premiums for a low-income, low-predicted spending, low-risk score male aged 65 to 70.[41] Higher income and higher risk (as captured by predicted spending) consumers are less sensitive to premiums, while older and riskier consumers are more responsive to coverage benefits. Conditional on predicted spending, consumers with higher risk scores are less responsive to benefits.[42] Gender does not meaningfully change consumers' responsiveness to premiums or benefits. The average post-subsidy premium elasticity is -0.9, conditional on plans with positive premiums. However, the regulatory environment restricts the ability of firms to increase pre-subsidy plan prices without offsetting changes to benefits. Due to this regulation and the extensive level of subsidization, the average demand elasticity with respect to pre-subsidy plan prices is an order of magnitude larger (-9.3, not in the table). I provide further details

[41] The discrepancy between my finding and those of Abaluck and Gruber (2011) are, in part, due to \$1 in benefits translating to less than a \$1 reduction in expected spending, which inflates the coefficient. I discuss this further in Appendix II.H.

[42] The difference between the role of predicted spending and risk scores in the demand estimates is potentially due to how risk scores compress the spending curve and rely on older data for risk assessment.

on how the regulation distorts firms perceived elasticity in Appendix II.I.[43] As I will show later, this elasticity implies reasonable markups for firms in this market.

Panel B of Table 1 presents consumers' preferences for fixed product attributes. Consumers have mixed preferences for dental benefits, prefer plans with more generous prescription drug coverage, and dislike those offering hearing aid benefits or vision coverage. Appendix Table 4 shows that every new Medicare generation has stronger preferences for MA, conditional on coverage, premiums, and all additional factors described thus far. Consumers who have attained higher degrees of education, have higher incomes, are riskier, or have employer-sponsored supplemental insurance are less likely to enroll in MA.

Panel C of Table 1 presents consumers' quality preferences under the assumptions of informed choice and linear quality preferences. The most valued quality category is Medical Outcomes, closely followed by Access to Care and Patient Experience. Process and Intermediate Outcomes quality—primarily associated with preventive care and chronic condition management—are the least valued. Therefore, the estimates indicate that consumers place great value on having good access to high-quality hospitals and physicians and place substantially less value on insurers facilitating preventive care or monitoring their health. In terms of yearly premiums, a low-income, low predicted spending, low-risk score male aged 60 to 75 would be willing to pay $4,036 a year to access the highest possible quality of Medical Outcome but only $1,614 for the highest Intermediate Outcome quality. Mapping the estimates to the data, consumers are willing to pay $12,004 for the median quality plan, which is about 24% more than what this plan charges to consumers and the government combined. This suggests that there are substantial gains from trade in the market. Therefore, we should not expect the optimal scoring policy to steer consumers away from the MA market segment.

VI.A.3 Informational losses: Consumers value quality, which they cannot observe. To quantify this inefficiency, I compute the surplus loss from consumers' incomplete information, holding product attributes fixed. The average consumer loses $199.3, or a third of a year's premiums, due to two informational frictions. First, *within scores*, the quality of products is indistinguishable. For example, the average spread in quality between the best and worst 4-star plans is equivalent to a $367.8 difference in premiums. Second, *across scores*, misalignment between consumers' preferences for quality categories and their rela-

[43]This is the elasticity relevant for firms' pricing decisions and, in particular, a single-product monopolist with constant marginal cost and no Part D coverage would set prices to meet an elasticity of -1. The table also displays premium elasticities comparable to those of Miller *et al.* (2022). Their estimate is -2.6 using similar data but a different model. In my model, this premium elasticity would imply excessive price elasticities and negligible firm markups.

Table 1: Demand Estimates

Panel A:	**Premium (α_i)**		**Benefits (β_i)**	
Mean preferences	-1.361***	(0.377)	3.090***	(0.498)
Medium income	0.001	(0.054)	0.116	(0.068)
High income	0.221***	(0.057)	0.036	(0.071)
Female	-0.063	(0.046)	-0.006	(0.058)
Age group < 65	-0.115	(0.086)	0.104	(0.108)
Age group $\in [70, 75)$	0.038	(0.060)	0.137**	(0.053)
Age group $\in [75, 85)$	-0.072	(0.068)	0.400***	(0.058)
Age group ≥ 85	-0.158	(0.112)	1.140***	(0.090)
Medium spending	0.110*	(0.053)	0.140***	(0.036)
High spending	0.149**	(0.056)	0.201***	(0.041)
Medium risk score	-0.023	(0.058)	-0.062	(0.071)
High risk score	0.072	(0.096)	-0.251*	(0.106)

Panel B: Other product attributes (λ^a)			**Panel C: Quality preferences (γ)**		
Dental cleaning	1.882***	(0.077)	Access	5.338***	(0.160)
Dental exam	-2.573***	(0.116)	Intermediate	2.198***	(0.096)
Dental x-ray	0.777***	(0.053)	Outcome	5.493***	(0.603)
Drug deductible	-0.001***	(0.000)	Patient	4.052***	(0.194)
Enhanced drug coverage	0.072***	(0.018)	Process	2.470***	(0.265)
Fluoride treatment	-0.536***	(0.031)			
Hearing aids	-0.332***	(0.041)			
Hearing aids fitting	-0.164***	(0.037)			
No part D coverage	-1.816***	(0.029)			
Vision coverage	-0.028	(0.031)			

N	36447	Log likelihood	-5.403	Mean premium elasticity ($p^C > 0$)	-0.968

Notes: This table presents the estimated consumer preferences, except demographic preferences (λ^l), which are shown in the Online Appendix Table 4. Panel A presents consumers' preferences for premiums and benefits, measured in thousands of dollars per year. The normalized group (captured in the mean preferences row) corresponds to low-income, low-predicted spending, and low-risk score males aged 65 to 70. Spending, risk score, and income groups are defined according to terciles of the population distribution across all years. Panel B shows consumers' preferences for additional product attributes. Except for Drug deductible, all variables indicate whether the plan offers coverage for the corresponding element (e.g., "Dental cleaning" stands for whether the plan offers coverage for dental cleanings). Panel C shows the estimated consumer preferences for each quality dimension. As quality is bounded in the unit interval, dividing these numbers by the absolute value of premium preferences results in consumers' willingness to pay for maximum quality in each dimension for a year, in thousands of dollars. Loglikelihood is adjusted by MCBS sampling weights. Standard errors are homoskedastic and corrected for multi-stage estimation using the delta method. *p<0.05, **p<0.01, ***p<0.001. Online Appendix Table 5 shows the effects of the instruments on these estimates.

tive contribution to the score makes it such that higher-scoring products can have lower quality-utility than lower-scoring ones. On average, 22.7% of plans have a lower-scoring alternative delivering higher quality-utility. Decomposing the surplus loss into these factors reveals that 94.5% stems from across-score frictions.[44] Within-score frictions are limited by firms' incentives to target the lower boundaries of scores.

VI.A.4 Selection: There is extensive literature studying selection into MA (Newhouse and McGuire, 2014; Brown *et al.*, 2014; McGuire and Newhouse, 2018). The demand model specified here enables an additional margin of selection based on attracting profitable consumers with quality (Glazer and McGuire, 2006; Glazer *et al.*, 2008). Appendix Figure 1a shows consumers' WTP for quality decreases in their risk scores.[45] As higher risk scores contribute more to a firm's profits due to risk adjustment, firms have an incentive to attract high-risk scores with lower quality. Therefore, risk adjustment contributes to the underprovision of quality in the market. Appendix Figure 1b, however, shows that the distortive effect of risk scores on the distribution of WTP for quality is small, suggesting that selection incentives are likely to play a secondary role in the overall results. Relatedly, and in light of recent evidence of insurers' upcoding (Geruso and Layton, 2020) and selection practices (Fioretti and Wang, 2021), Appendix II.J provides evidence that manipulation and selection are unlikely to play an important role in the scoring design problem.

VI.B Supply

VI.B.1 Insurance marginal costs: Insurers' pricing first-order optimality condition equates marginal revenue with marginal costs.[46] Since revenue depends only on observed demands, prices, and estimated elasticities, this condition can be used to recover the marginal cost parameters ($\boldsymbol{\theta}^c$). Assuming marginal costs are linear, the resulting condition is

$$\underbrace{\boldsymbol{p}_f + R(\boldsymbol{p}_f, \boldsymbol{z}_f)}_{\text{revenue per consumer}} + \underbrace{(\nabla \tilde{\boldsymbol{D}}_f')^{-1}(I + \nabla R_f(\boldsymbol{p}_f, \boldsymbol{z}_f))\tilde{\boldsymbol{D}}_f}_{-\text{profit margin}} = \underbrace{\boldsymbol{\theta}_q^{c\prime}\boldsymbol{q}_f + \boldsymbol{\theta}_a^{c\prime}\boldsymbol{a}_f + \boldsymbol{c}_f}_{\text{marginal cost} = \boldsymbol{C}(\boldsymbol{q}_f, \boldsymbol{a}_f, \boldsymbol{\theta}^c)}, \tag{12}$$

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. On the right-hand side, I have decomposed the firm's marginal cost into its

[44] This is done by simulating a scenario without within-score frictions: Consumers first choose a plan based on expectations and then get to adjust their choice among plans of the same score with full information.

[45] This correlation is driven not by the small effect of risk scores on the premium parameter but by the correlation between risk score and income and predicted spending.

[46] As the firm's problem is not differentiable at the regulatory kink (benchmark), the FOC is only valid for prices away from this cutoff. However, in the data, no firm violates this condition.

quality components ($\boldsymbol{q}_f$), systematic observable components ($\boldsymbol{a}_f$), and residual ($\boldsymbol{c}_f$).

Variations in demand and regulation identify marginal costs. I assume that the residual cost variation in $\boldsymbol{q}_f$ conditional on contract identity and year is unsystematic and unknown when firms choose quality. Further, I assume that risk adjustment is perfect and heterogeneity across products and firms is large enough such that the observed prices form a locally stable equilibrium. Thus, marginal changes in demand or regulation do not discretely change equilibrium play and allow the identification of the marginal cost components.

Panel A of Table 2 presents the estimates of $\boldsymbol{\theta}_q^c$ when $\boldsymbol{a}_f$ includes contract, year, and market fixed effects and controls for bundled services. Quality's effect on marginal costs is identified by the residual correlation between marginal revenue and quality after accounting for market and national quality trends. The estimates indicate that a marginal improvement in Access and Outcome quality increases the marginal cost of insuring a unit-risk consumer by \$30 and \$15 per month, respectively. As both categories are improved by changing provider networks, those costs likely reflect higher prices from marginal providers. A marginal improvement in Intermediate quality entails additional monitoring and maintenance of chronic conditions, resulting in a marginal increase in costs of \$104 per month. In contrast, improvements in Process and Patient quality lower marginal costs by \$176 and \$215 per month, respectively. For Process, this is likely due to its effect on preventive care and managing expensive chronic illnesses, which can prevent costly hospitalization (Newhouse and McGuire, 2014). Having better physicians in the network (i.e., Patient quality) is likely associated with similar improvements and might make patients more likely to adhere to preventive and diagnostic care. Nevertheless, these improvements might come at the expense of significant investment costs.

These estimates imply reasonable markups for insurers, with an average of 10.5%. Using claims data for the top insurers during 2010, Curto *et al.* (2019) estimated an average cost of \$590 per enrollee risk-month in medical costs, or \$680 in adjusted 2015 dollars. My estimate for the same set of firms is \$758, including administrative costs. This comparison suggests that about 11% of marginal cost is administrative, which is consistent with the level of involvement of MA insurers with their enrollee's health.

VI.B.2 Investment costs: As quality investments are subject to risk, observed quality realizations might differ from their targets and violate insurers' investment first-order optimality condition. The first step in adapting the identification and estimation strategy to this challenge involves recovering the investment risk distribution. I assume that for any contract c, its realized quality and investments in a dimension k are related by the map-

Table 2: Quality's Insurance Costs, Investment Costs, and Marginal Welfare

Term	Access	Intermediate	Outcome	Patient	Process
Panel A: Insurance cost					
Linear ($\boldsymbol{\theta}_q^c$)	30.077	104.038***	15.903***	-215.450***	-176.811***
	(16.928)	(12.858)	(3.866)	(57.830)	(27.936)
Panel B: Investment cost					
Common linear ($\boldsymbol{\mu}_k$)	1.514***	-0.004	0.545**	-2.595***	3.010***
	(0.177)	(0.182)	(0.208)	(0.480)	(0.374)
PCP rate ($\bar{\boldsymbol{\mu}}_k$)	-0.006***	-0.005***	-0.013***	0.002	-0.007**
	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)
Quadratic ($\boldsymbol{\mu}'_k$)	-1.366***	1.809***	3.071***	9.109***	-0.475
	(0.197)	(0.312)	(0.472)	(0.0.614)	(0.471)
Panel C: Marginal Welfare					
Per contract-year	62.376	17.647	84.900	65.287	65.225
	[3.9, 73.3]	[1.6, 25.1]	[8.4, 107.6]	[0.7, 102.3]	[2.5, 78.4]

Notes: Panel A presents the marginal insurance cost coefficients associated with plan quality ($\boldsymbol{\theta}_q^c$) in dollars per unit-risk member and month. The regression includes controls for additional contract benefits ($\boldsymbol{\theta}_a^c$), which are shown in Online Appendix Table 6, and fixed effect controls for plan types (HMO, PPO, Regional plans, and PFFS), county, year, and contract identifiers. Standard errors in parentheses are heteroskedasticity robust. $N = 28,966$, $R^2 = 0.529$. Panel B presents the estimated investment cost parameters in millions of dollars per hundred thousand Medicare beneficiaries annually. The regression includes spillover cost components ($\mu''_{k,k'}$), shown in Appendix Table 7, and linear cost terms for the interaction between the top six firms' identities, and each quality dimension, which are omitted for space. The mean PCP rate is 83.1 per hundred thousand individuals. $N = 7,684$. Panel C shows the average derivative of total welfare ($TW(\psi, 1, 1)$ in equation (7)) with respect to each contract's quality in each dimension, in millions of dollars per year. The interquartile range is provided in brackets. *p<0.05, **p<0.01, ***p<0.001.

ping $q_{ck} = \Phi_k(x_{ck} + \epsilon_{ck})$ where $\Phi_k(\cdot)$ is an increasing function mapping the real line to the domain of quality dimension k and ϵ_{ck} is the investment shock.[47] Under certain regularity conditions, the distribution of ϵ_{ck} can be estimated non-parametrically using standard deconvolution results given observed quality realizations (Schennach, 2016). Appendix II.K details the formal identification and nonparametric estimation procedure for risk.

Given the identified distributions of investment risk and quality, we can evaluate firms' in-

[47]The definition of Φ_k is arbitrary as x_{ck} is a modeling device rather than a true investment metric. In practice, I take $\Phi_k(x) = \Phi(x)(\bar{q}_k - \underline{q}_k)$ where $\Phi(\cdot)$ is the standard normal CDF and $\bar{q}_k, \underline{q}_k$ are the upper and lower bound of the support of quality dimensions k in the data. In practice, the upper bound equals 1 for all categories, but the lower bound is constrained above zero by minimum quality regulation.

vestment optimality conditions in expectations. Formally, the first-order condition of investment for firm f in category k in year t equates marginal revenue ($\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}]$) with marginal investment cost ($\frac{\partial I_f(\boldsymbol{x}_{ft})}{\partial x_{ckt}}$). Given observed quality $\boldsymbol{q}_{ft}$, we can decompose the marginal revenue into its conditional mean and variance, resulting in the condition:[48]

$$\mathbb{E}[\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}]|\boldsymbol{q}_{ft}] = \frac{\partial I_f(\boldsymbol{x}_{ft})}{\partial x_{ckt}} + \nu_{ckt} \qquad \mathbb{E}[\nu_{ckt}|\boldsymbol{q}_{ft}] = 0 \tag{13}$$

The first term corresponds to the posterior expectation of marginal insurance profits given observed quality, the second to marginal investment cost, and the third to the conditional variance of marginal profits. As noted by the second equality condition, equation (13) is a regression equation. To operationalize it, I model firms' investment costs as

$$I_f(\boldsymbol{x}_{ft}) = \sum_{c\in\mathcal{C}_{ft}, k\in\mathcal{K}} \underbrace{M_{ct}}_{\text{population}} \Big(\underbrace{\mu_{ckt}\tilde{x}_{ckt} + \frac{\mu'_k}{2}\tilde{x}^2_{ckt}}_{\text{category-specific cost}} + \underbrace{\sum_{k'\neq k}\frac{\mu''_{k,k'}}{2}\tilde{x}_{ckt}\tilde{x}_{ck't}}_{\text{cross-category spillovers}}\Big) \tag{14}$$

Where $\tilde{x}_{ckt} = x_{ckt} - \underline{x}_{kt}$ and $\underline{x}_{kt}$ is the lowest level of investment a firm can deliver to participate in a state. Anything above this level requires forming a network or writing contracts to promote quality. Above, M_{ct} denotes the total Medicare-eligible population across counties where contract c is offered, measured in hundreds of thousands. The first two terms within the parenthesis are the category-specific quadratic costs, where I parametrize the first as $\mu_{ckt} = \mu_k + \tilde{\mu}_{k,f(c)} + \bar{\mu}_k PCP_{ct} + \epsilon^{\mu}_{ckt}$. In this expression, the first coefficient captures common investment costs, while the second captures firms' cost advantages. The third coefficient, $\bar{\mu}_k$, captures how the cost of dimension k depends on essential inputs, namely the local availability of primary care physicians (PCP) in the counties of operation of contract c. Finally, ϵ^{μ}_{ckt} is an iid mean-zero unobserved shock to investment costs. The final term in Equation (14) captures cross-category spillovers in investment, which I assume are symmetric. This captures, for example, how improving chronic condition management (Intermediate) can reduce the cost of improving medical outcomes (Outcome) or vice versa.

I use the optimality condition of Equation (13) and the investment cost function of Equation (14) to evaluate the implied moment condition ($\mathbb{E}[\nu_{ckt}|\boldsymbol{q}_{ft}] = 0$) using observed quality. This involves replacing $\tilde{x}_{ckt}$ in (14) with its closest observable analog, $\Phi_k^{-1}(q_{ckt}) - \Phi_k^{-1}(\underline{q}_{kt})$, where $\underline{q}_{kt}$ is the minimum quality for dimension k in year t. This replacement and the unobserved cost shock ϵ^{μ}_{ckt} create an endogeneity problem: The realized quality will tend to be

[48] Appendix II.L describes how I estimate firms' rational expectations about rivals' actions, which are necessary to evaluate this expectation. Appendix II.M shows that the left-hand side of this expression is a function of only identified distributions and has an analytical expression.

greater in years when it was cheaper to produce and when investment shocks were larger. To address this, I use three instruments based on the scoring design variation, consumers' unobserved preferences, and local factors affecting the need for quality investments.

The first instrument corresponds to the product of category k's contribution to the rating in year t and the inverse of consumers'' unobserved preferences for contract c.[49] Both greater category contribution and lower consumer preferences tend to push firms to increase investments. The second and third instruments correspond to the product of the category contribution with an index for the availability of healthy foods in each county and the share of a county's population older than 65. Both factors indicate counties that are more vulnerable and where investments might be more impactful. The instruments vary across time and contracts due to design variation, shifting preferences, and differences in counties of operation. The exclusion restriction assumes unsystematic cost shocks are orthogonal to policy changes and consumers' unobserved plan preferences.

I estimate the investment cost parameters using GMM. In addition to the moment condition formed by the instruments and the expected optimality condition, I include two moments based on the reported total investment per contract in 2015: one matching observed and predicted investments in levels and one as a share of insurance profits. To avoid mixing contracts with distinct cost structures, I limit attention to HMO and PPO contracts.[50]

Panel B of Table 2 shows firms' common linear and quadratic cost terms. The estimates show that Access quality cost is concave, suggesting economies of scale in expanding provider networks and facilitating appointments. In contrast, Intermediate, Outcome, and Patient quality costs are convex. A rationale for the first two is that as higher-quality providers are brought into the network to improve performance, the leverage of marginal providers increases, allowing them to extract more of the insurer's profits. Patient experience quality is particularly costly to modify, likely because there is no direct investment that allows insurers to alter patient satisfaction. Process quality is linear in cost, likely because it consists of paying for simple procedures like lab work and screenings. The higher availability of primary care physicians reduces the cost of quality across all dimensions except for patient experience. More physicians reduce the cost of expanding networks and the leverage of marginal providers, which is likely associated with more competition across providers of screening and labs. It is also reasonable that it does not affect patient satisfaction, as the

[49]Formally, the instrument is $\omega_{kt}(\frac{1}{|\mathcal{J}_{ct}|}\sum_{j\in\mathcal{J}_{ct}}\tilde{\xi}_{jt})^{-1}$ where $\tilde{\xi}_{jt}$ is the average across counties of the demand residual preferences $\tilde{\xi}_{jmt}$.

[50]HMO and PPO account for 81% of enrollment. This excludes PFFS contracts, which do not form networks, and Regional PPO contracts, which have broad networks that often cross multiple state lines.

additional PCP might not be better than those found in tighter markets.

Appendix Table 7 shows the spillover terms. Almost all effects are negative, indicating that investing in one dimension reduces the cost of investing in others. Investing in Intermediate or Process quality vastly reduces the cost of improving patient experience. Plausibly, better monitoring and diagnostic care improves physicians' information about patients, improving patients' experience when seeking care. Another substantial spillover is found between the Outcome and Intermediate Outcome categories, likely due to the detrimental effect that deteriorating chronic conditions have on medical outcomes overall.

The investment cost estimates indicate that the median contract invests 12% of its insurance profits back into quality. For the available data, the true median for 2015 is 15%, while the predicted value for the same set of contracts is 19%. Despite some negative cost coefficients, the marginal cost of increasing quality is positive for all firms in all quality dimensions, even considering the effects on insurance marginal cost. It is worth noting, however, that the estimated cost function does not include the fixed cost of participating in the market. The investment cost structure of Equation (14) is normalized such that firms investing at the minimum level ($\underline{x}_{kt}$) pay an investment cost of zero.

VI.B.3 Efficiency: I evaluate the efficiency of quality provision by computing the marginal welfare value of quality (i.e., $\frac{\partial TW(\psi,1,1)}{\partial x_{ckt}}$), holding prices fixed. Panel C of Table 2 shows that for the average contract, a marginal increase in any dimension would increase consumers' surplus by more than it would cost to produce. The most underprovided dimension is Outcome quality, with a marginal value of $84.9 million per year, and the least is Intermediate quality, with a marginal value of $17.6 million. Appendix Table 8 shows the results of regressing the marginal welfare value of quality on HHI, the category's contribution in the soring design, and category-contract fixed effects. More concentrated markets and categories with lower contributions are associated with larger derivatives and, thus, greater underprovision. This is consistent with the Spencian distortion and scores' ability to influence it. The following section revisits the regulator's problem and examines how optimal scoring policies might address these inefficiencies in quality provision.

VII Scoring Design

In this section, I solve the optimal scoring design problem within the monotone partitional class and decompose its regulatory mechanisms. I use the results to explore the effects of asymmetric information, moral hazard, and regulatory bias. Additional results regarding preference heterogeneity and competition are presented in Appendix III.

VII.A Approach

The designer seeks to maximize total welfare, $TW(\psi, \rho^F, \rho^G)$ in Equation (7), by choosing a scoring rule ψ from a class Ψ. The designer recognizes the effect of its policy on equilibrium investments, prices, beliefs, and enrollment choices. This endogenous investment response presents the designer with a trade-off: For any fixed investment distribution, more information helps consumers choose and might make competition more effective. However, firms might invest inefficiently under full information—a distortion coarser information can regulate. Therefore, the scoring rule must trade off information for efficiency.

Solving this trade-off is challenging. First, scoring rules are discontinuous mappings from quality space down to a few scalars. There are no known optimality conditions, and a priori, the loss from approximations is unbounded. Second, because the regulator computes an expectation over quality, evaluating designs requires integrating over a continuum of counterfactual subgame equilibria. I draw on two insights to address these challenges and develop a method to divide the design problem into a series of smaller, manageable ones.

First, Appendix III.A shows that any monotone partitional score is a composition of a polynomial *aggregator*, aggregating multidimensional quality into an index, and a *cutoff* function, which partitions the index into scores. Therefore, we can find the optimal design by finding the best one for all subproblems constrained to a particular number of cutoffs and aggregator polynomial order (i.e., the boundary curvature). Each of these problems is moderately simple, conditional on being able to compute the regulator's integral.

The second insight addresses the integral and comes from Aumann *et al.* (1995), who note that selecting a disclosure policy is akin to choosing a distribution of posterior beliefs. In scoring design, the analogous statement is that each score generates a distribution over qualities, score valuations ($\mathcal{E}[\boldsymbol{\gamma}'\boldsymbol{q}|r,\psi]$), and marginal quality costs ($\boldsymbol{\theta}^{c\prime}\boldsymbol{q}$). This observation enables a strategy that first evaluates the objective over a large collection of potential outcomes and then associates each score with a distribution over these evaluations. Therefore, the integral of any policy is a known weighted sum of points in the grid.

As scoring policies are discontinuous and investment competition occurs over multiple dimensions, the uniqueness and existence of equilibria are challenging to guarantee. For any conjectured policy, I find the game's equilibrium by intersecting firms' best responses starting from the status quo. This Gauss-Seidel approach ensures that if convergence is attained, it is to a unique Bayes-Nash solution that is nearest in the best-response distance. This also ensures that convergence would fail if such equilibrium does not exist or is not unique. A secondary problem in solving the optimal scoring design is that the regulator's objective

has many local optima. To find the global maximum, I use the algorithm of Malherbe and Vayatis (2017), which provides convergence guarantees for Lischitz continuous functions of multiple bounded arguments. Appendix III.B offers further details on implementation.

The following results focus on the markets included in the MCBS 2015 data, covering nearly 22 million beneficiaries. Except for subsection VII.G, all the results are under the *informed choice* assumption. Given mixed evidence on the extent of selection into MA (Newhouse *et al.*, 2015), I omit subsidy spending from the main analysis, focusing on the case of $\rho^F = 1$ and $\rho^G = 0$. I discuss different objectives in subsection VII.E.

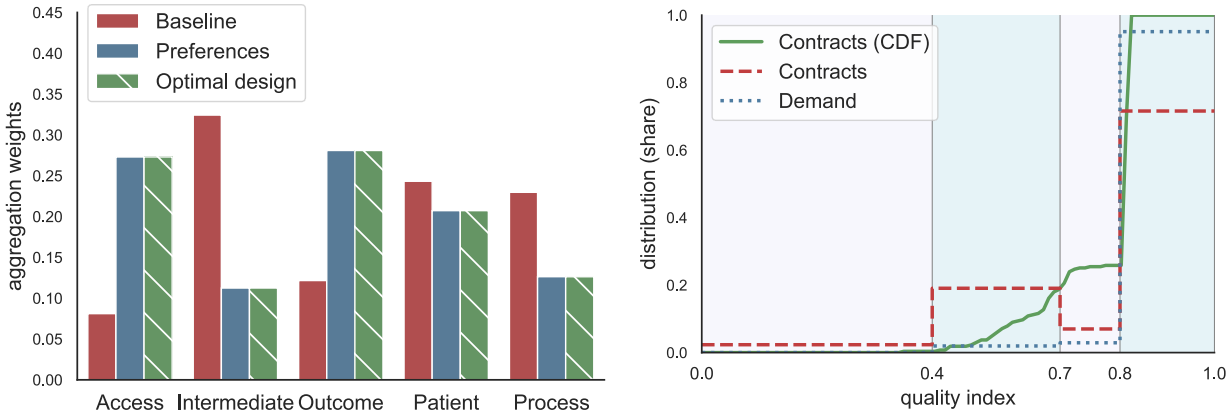
The first row of Table 3 shows the model *baseline*, or status quo, in 2015. The second row provides the market status and welfare changes under a counterfactual of full information to help benchmark the following results. However, welfare numbers derived under this scenario must be taken cautiously as they might entail substantial losses from increased choice complexity, which is unaccounted for in this analysis.

VII.B Optimal design

Figure 5 shows the optimal monotone partitional design, and the first row of Panel A in Table 3 shows its effects on the market. This solution was constrained to using at most fifteen partitions and a quadratic aggregator. The optimum, however, features a linear aggregator and four scores, indicating that the constraints are not binding. The optimal policy has three key features: the lowest score pools quality at the bottom of the distribution; the cutoff function has limited granularity, using only four scores; and the aggregator is aligned with consumers' preferences. Next, I discuss these features and their key mechanisms.

VII.B.1 Pooling at the bottom: The first stark difference between the new design and the Star Ratings is how they classify low-quality plans. The Star Ratings partition the quality space uniformly, allowing consumers to distinguish between low and medium-quality plans. This information is valuable to consumers, yet it distorts quality provision. By pooling at the bottom, the new design uses within-score informational frictions to induce low posterior beliefs among consumers for low-scoring plans; this shifts their demand toward better scores, incentivizing investment and remedying the quality underprovision problem.

Figure 5b shows that contract quality is concentrated above the last scoring threshold. As 71.5% of all contracts fall within this score, the average consumer in the counterfactual chooses among higher quality products with greater information about them. As a result, top-scoring contracts enroll over 95% of all MA consumers. In contrast, there is limited offering and demand for products at the bottom of the distribution. If one applied the



(a) Aggregation weights

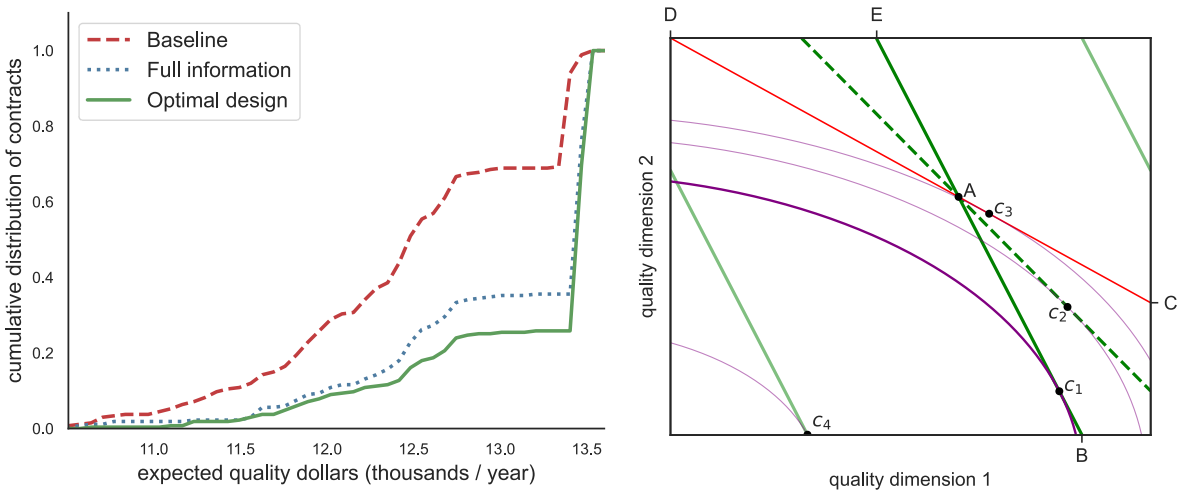
(b) Quality distribution and cutoffs

Figure 5: Optimal Design

Notes: The optimal design comprises aggregation weights that reduce multidimensional quality into a single index and cutoffs that partition the index into scoring regions. Figure (a) compares the optimal aggregation weights (green) with CMS's average aggregator for 2015 (red) and consumers' preferences normalized to a unit sum (blue). Figure (b) shows the scoring cutoff along the quality index, with segments indicating different scores. The green line shows the equilibrium cumulative distribution of contracts across scores, the red, the share of contracts per score, and the blue, the share of demand per score. Quality bunches to the right of the last cutoff as firms have no incentives to invest beyond this point. Bunching is right-shifted due to investment risk and a (mostly) concave demand function, which induces risk attitudes in firms.

optimal design to the baseline quality (not shown), 6% would be classified as the lowest score and 51% as the second-lowest. As the equilibrium demand for these contracts is a meager 0.004% and 1.9%, respectively, the new design incentivizes insurers to invest. In equilibrium, only 2% of contracts fall within the last score and 19% within the second-lowest. Contracts at the bottom score are virtually exiting the market, with investments matching the minimum standard. Overall, the new design leverages the same mechanism as illustrated in Section II. Figure 6a shows that the left tail of the quality distribution in the regulated market is shifted inward relative to a full-information counterfactual, matching the model's predictions and highlighting the alleviation of underprovision.

VII.B.2 Limited granularity: The granularity of scores equals the number of potential investments firms might consider optimal. Intuitively, firms aim at the cutoffs since interior investments do not translate to increased demand. This observation is known as the *delegation equivalence* of scores (Kolotilin and Zapechelnyuk, 2019), which explains how scoring granularity affects the supply of heterogeneous products. The counterpart to this effect is consumers' heterogeneity in WTP for quality. As preference heterogeneity grows, so does the optimal variety of products. Hence, a more granular scoring system allows a larger



(a) Quality underprovision

(b) Weight alignments

Figure 6: Scoring design mechanisms

Notes: Figure (a) plots the distribution of quality in the baseline, full-information benchmark, and optimal design. Quality is measured according to consumers' WTP, in thousands of dollars per year. The distribution does not match that of Figure 5 due to differences in the x-axis. Figure (b) illustrates the aggregation mechanism with two quality dimensions and four scores. Line EB is the cutoff separating the second from the third score, and DC is the consumers' indifference curve. The misclassification region is $DEA + ABC$, since consumers prefer products in DEA over those ise ABC. The dashed green line represents a potential redesign. Purple concave lines are a firm's isocost curve under different total investment levels.

variety of products to match with consumers of different tastes. The trade-off, however, is that firms' incentives to provide quality suffer from the Spencian distortion, and as their production flexibility grows, these distortions increase. Hence, there is only one optimal cutoff in a setting of homogeneous preferences and firms since there is a unique optimal quality. In contrast, the optimal granularity is infinite in an environment with heterogeneous consumers and firms but no distortions. In MA, the optimal granularity is 4, five fewer than the status quo. Any more, and the loss from quality distortion exceeds the gains from variety.

VII.B.3 Aggregation weights: The final feature of the new design is its aggregation weights. In quality space, these weights determine the slope of boundaries separating one score from the next. Figure 6b illustrates this in a two-dimensional case, with line BE being the boundary between the second and third score and DC being the consumers' indifference curve. The new design aligns scoring boundaries with consumers' indifference curves, rotating BE to match DC. This change ameliorates two failures caused by quality aggregation.

The first loss stems from the multitasking moral hazard problem noted in Section V.B. Ignoring investment risk, firms first choose which score their plans should have and then

find the cost-minimizing way to attain such a score. For example, in Figure 6b, point c_1 marks the tangency of a firm's isocost curve (purple) with the scoring threshold (green), which would be the efficient investment combination for it to attain the third score. For the regulator, however, this decision introduces a multitasking moral hazard problem as firms ignore consumers' preferences over the relative allocation of quality. Aligning boundaries and preferences eliminates this problem by rendering firms' incentives to substitute investments across quality dimensions similar to consumers' marginal rate of substitution.

The second loss from aggregation is the across-scores informational distortion, visible in Figure 6b since products in the triangle ΔDEA are preferred by consumers to the higher-scoring ones in the quadrilateral formed by points A, B, C, and the bottom-right corner. The optimal alignment of boundaries and preferences eliminates the misclassification region. As shown in Table 3, the mean squared error of consumers' beliefs regarding plan quality drops by 75.3%. The informational gains from eliminating misclassification are slightly offset by increased within-score informational frictions imposed by pooling at the bottom.

There are no guarantees that aligning aggregation weights and consumers' preferences is optimal. The design must account for cost heterogeneity across firms as changes in alignment force firms to adjust their overall investment. Discontinuities in demand across scores imply that marginal changes in cost might result in a discrete change in firms' optimal investment strategies. For example, in Figure 6b, as EB rotates around point A, the firm that formerly invested in c_1 must now invest in a higher level c_2 to obtain the same score. As investment costs are mostly convex, further tilting might dissuade the firm from maintaining a score of 3 at point c_3, pushing it to a score of 2 at c_4 and a substantially lower quality. This trade-off between alignment and incentivizing production from heterogeneous firms is more noticeable in the optimal designs for other regulatory objectives, as discussed below.

VII.C Welfare

The third row of Table 3 shows the estimated welfare gains from replacing the MA Stars with the optimal design. Per Medicare beneficiary, the alternative increases consumer surplus by \$47.99, slightly more than an average monthly premium payment in the baseline. Firm profits increase by \$107.75 per beneficiary or 24.6% per MA enrollee (baseline value not in table). The change in scores induces a significant change in investment, with the new equilibrium investment being nearly three times as high per contract as in the baseline. An examination of the changes (see Appendix Figure 2) reveals that this is driven by contracts that obtain between 2 and 3.5 stars in the baseline and that are predicted to invest enough to obtain the highest score in the counterfactual. Among those plans, quality is increasing

Table 3: Scoring Design Equilibrium Impacts

	Insurance Market Outcomes								Welfare Change			Compensating Var.	
	Premium	Benefits	Insurance Markup	Investment Cost	Contract Quality	Beliefs MSE	MA share	Subsidy Spending	Δ Consumer Surplus	Δ Firm profits	Δ Total welfare	Δ Info.	Δ Quality
Baseline	42.13	82.60	13.2%	5.70	15.73	0.94	29.4%	9.84					
Full Info.	44.24	79.13	16.1%	14.48	16.36	0.00	31.6%	9.84	66.14	66.98	133.12	28.90	65.71
Panel A: Alternative Designs Under Informed Choice													
Optimal	49.28	77.75	17.0%	15.27	16.39	0.23	33.5%	9.85	47.99	107.75	155.74	70.45	90.14
Certification	49.80	77.80	17.0%	15.93	16.42	0.24	33.6%	9.85	48.71	104.25	152.95	68.23	105.24
CMS-Cert.	39.86	80.37	15.7%	14.34	16.26	0.10	30.7%	9.83	63.82	48.12	111.94	27.13	64.86
CMS-Full	40.40	79.73	15.6%	14.28	16.32	0.09	30.8%	9.83	69.13	46.66	115.80	19.51	62.31
Top-Revealing	49.67	77.37	17.0%	14.05	16.25	0.21	34.0%	9.85	51.92	110.79	162.71	73.02	113.56
Panel B: Optimal Certification Under Private Cost Types													
Coarse	49.80	77.80	17.0%	15.78	16.40	0.25	33.5%	9.85	48.99	104.47	153.46	70.36	105.54
Top-Revealing	49.80	77.80	17.0%	14.35	16.23	0.30	33.7%	9.85	53.00	106.33	159.32	63.35	105.42
Panel C: Optimal Design Under Alternative Objectives													
$\rho^F = \rho^G = 0$	37.78	81.93	14.9%	14.36	16.33	0.30	29.9%	9.83	75.26	25.37	100.63	-3.21	51.63
$\rho^F = 0.5, \rho^G = 0$	46.92	78.24	16.5%	15.39	16.39	0.08	32.8%	9.85	59.01	88.15	147.17	58.30	75.80
$\rho^F = \rho^G = 1$	48.23	77.65	16.9%	14.70	16.37	0.18	33.4%	9.84	47.10	105.67	152.77	67.32	103.27
Panel D: Optimal Design Without Informational Impacts													
Optimal	49.94	77.80	17.0%	12.70	16.18	0.22	34.3%	9.86	62.71	103.49	166.20	0	91.67

Notes: This table presents the changes induced by alternative designs on the market. The first row corresponds to the baseline values, and the second to a full information counterfactual. Panel A presents the results for the optimal coarse monotone partitional design, the optimal certification design, the optimal certification subject to CMS's weighting scheme, the optimal multi-scored design subject to CMS's weighting scheme, and the optimal top-revealing design (which is not a coarse monotone partitional design). Panel B shows the results assuming that insurers have private information about their investment cost types, subject to either a simple certification scheme (coarse) or a certification scheme that fully reveals quality at the top (Top-Revealing). The welfare numbers for this panel integrate over the regulator's cost uncertainty and thus are not directly comparable with those of other panels. Panel C shows the results under alternative regulatory objectives. All dollar values are adjusted to 2015 dollars using medical CPI and weighted by enrollment. Panel D shows the optimal design if consumers' interpretation of the scores is invariant to the design and fixed to the status-quo value. Premiums and benefits are measured in monthly dollars, investment costs in millions per contract year, and government subsidies in thousands per member year. The latter includes only FFS spending on TM and benchmark plus rebate subsidies on MA. Contract quality corresponds to the mean realized true quality per contract, measured annually in thousands of premium-equivalent dollars. Belief MSE is the mean squared error between consumers' expected quality utility from choosing each contract and its true quality utility. MA share corresponds to the total market share of MA among all Medicare beneficiaries. These numbers are not adjusted for potential differential selection into MA; see Appendix III for adjusted numbers.

by as much as 16.6%, while the overall average change is 4% as shown in Table 3. This increase in spending is partially offset by a 17% increase in premiums and a 6% decrease in benefits, translating to a 3.8 percentage-point increase in insurance markups.

The new design also increases the predicted enrollment share of MA by 4.1 percentage points. Consumers switch from TM to MA as quality and information improve, allowing them to benefit from MA's generous cost-sharing. Consumers who switch to MA often choose plans that cost more to subsidize than TM, increasing subsidy spending by about ten dollars per beneficiary. This increase does not account for potential positive selection into MA nor changes in part D subsidies, which is discussed in Appendix III.D.[51]

VII.C.1 Asymmetric information and moral hazard: The alternative design changes information, quality, and prices. It alleviates frictions due to asymmetric information and firms' moral hazard and changes the degree of differentiation across firms, which affects market power over prices. To assess the value of these different channels, I compute the compensating variation associated with reverting either the informational structure or contract quality to its baseline value. Letting $(\boldsymbol{x}^*, \boldsymbol{p}^*, \Psi^*)$ denote the optimal investment, prices, and scoring policy, and $(\boldsymbol{x}^0, \boldsymbol{p}^0, \Psi^0)$ the baseline, the compensating variation value of quality is $CV_q = TW(\boldsymbol{x}^*, \boldsymbol{p}^*, \Psi^*) - TW(\boldsymbol{x}^0, \boldsymbol{p}^*, \Psi^*)$. The compensating variation value of information is computed analogously, replacing the role of investment targets and the scoring policy.

The two final columns of the third row of Table 3 show the estimated compensating variations. The regulator would have to distribute \$70.45 per Medicare beneficiary across market participants to offset the loss imposed by reverting information to its original state. The value of the new informational structure stems from the substantial reduction in choice frictions, as indicated by the sharp decline in consumers' beliefs' mean-squared errors. To offset the loss from reverting quality changes, the regulator would have to distribute \$90.14 per Medicare beneficiary across agents. The value of additional quality is driven solely by consumers' preferences. The sum of the two compensating variations exceeds the total welfare change as the compensating value of prices is negative.

Most of the welfare gains of this new scoring design stem from its role as a quality regulation policy. Regulating firms' moral hazard and offsetting the Spencian distortion contributes more to welfare than ameliorating informational frictions. In other words, the current regulatory environment is less effective at inducing quality than facilitating choice.

[51] For comparison, the predicted baseline number for 2015 is \$9835, while the true subsidy spending for this segment was \$10,581. The small difference is largely due to the restriction to MCBS counties with at least one HMO or PPO plan, which under-represents non-urban communities.

A key finding of this paper is that both targets can be improved using information alone. Moreover, consumers and firms would benefit from the change: Consumers from access to better insurance plans under better information, and firms from the coordination effect induced by the scores, which leads to market expansion and higher markups.

Table 3 also reveals that welfare under the optimal design exceeds that of full information, which has two implications for policy design. First, the ability to approximate full-information outcomes with simple coarse scores is valuable in settings where the underlying data are complex or subject to privacy regulations. For example, regulators might be unwilling to disclose the performance of small insurers since others might use it to identify their populations and discriminate against them. Yet, as in the example of Section II, consumers in a scored market can behave as if fully informed, even if they cannot detect large deviations in quality. Second, the gap implies that consumers still benefit from coarse information even if they are highly sophisticated Bayesian agents. Thus, the optimal design need not conflict with behavioral concerns about the ability of enrollees to process complex information—it is not the case that sophisticated consumers prefer complex signals of quality.

VII.C.2 Decomposition by design feature: The new design limits the scoring granularity and aligns the aggregation weights with consumer preferences. To isolate the different changes, I solve a series of constraint optimal design problems that gradually incorporate these features.[52] First, I find the optimal certification scheme that preserves CMS's average aggregation weight for 2015. The new design—whose equilibrium impact is shown on the third row of Panel A in Table 3—incorporates only the effect of pooling quality at the bottom. It attains 71.8% of the welfare gains of the optimal design, largely through the inducement of higher quality at lower premiums. The resulting design is approximate to certifying only if contracts exceed the 4.5-star threshold in the status quo. Allowing for additional scores while holding CMS's aggregator fixed results in the design described in the fourth row of Panel A in Table 3. This design features five scores and attains 74% of the welfare gains of the optimal design. This small improvement is due to additional offerings of lower-quality contracts for low-WTP consumers.

To isolate the effects of optimal weighting, I compute the optimal quality certification, shown in the second row of Panel A in Table 3. A simple but optimized certification is predicted to achieve 98.2% of the optimal design's welfare. This design addresses the informational loss from misclassification, the multitasking moral hazard problem, and the aggregate underprovision of quality on average. It fails only at incentivizing heterogeneous production.

[52]The resulting design is shown in Appendix Figure 3.

However, as low-WTP consumers have a free and high-value outside option, the loss from eliminating variety at the bottom of the distribution of quality is small. Accordingly, the certification cutoff is nearly identical to the highest cutoff of the optimal design.

VII.D Private information and connection to the theory

The analyses above assume the regulator knows insurers' cost structure, as identified from the data. This data, however, is only available after the market is realized and relies, in part, on the regulator's experimentation. Therefore, knowing how uncertainty about insurers' costs affects the optimal design is relevant. Solving this problem also closes the gap between this article's setup and that of its closest theoretical counterpart (Zapechelnyuk, 2020), which considers regulatory uncertainty over costs.

I model the regulator's uncertainty as five independent normal distributions over the heterogenous linear investment cost terms (μ_{ckt}), one for each quality category. I set the mean of the distributions to zero and the standard deviation equal to half the empirical standard deviation across firms in the estimated cost types. Embedding uncertainty into the optimal design problem substantially increases the computational cost of exploring alternative scores. To alleviate this, I focus exclusively on optimal certification designs, which are near-optimal in the main analysis, and take ten draws from the cost distributions. The first row of Panel B in Table 3 shows the resulting outcomes under the label "Coarse", while Appendix Figure 4 shows the resulting weights and cutoffs.

The results show that optimal certification in this setup is virtually identical to the one under known costs. Aggregation weights and cutoff are the same up to the first decimal. The welfare values shown in Table 3 integrate over cost uncertainty and thus are not directly comparable to those for previous results. The small improvement relative to the main analysis is due to the certification filter minimizing the exposure of consumers to bad cost types and low quality and increasing the reward low-cost firms can derive from the market.

Zapechelnyuk (2020) proves that the optimal design for a stylized monopolistic case is a *top-revealing* certification policy: All qualities below a threshold are pooled together, while those above are fully revealed.[53] This class of designs is not nested within the coarse monotone partitional designs over which the main solver searches. However, modifying the exploration strategy to optimize among top-revealing policies is simple. Appendix Figure 4 shows the optimal top-revealing certification, and the second row of Panel B in Table 3 shows

[53] Zapechelnyuk (2020) considers consumer surplus-maximizing designs. The same proof strategy can be used to show that the welfare-maximizing design also consists of top revelation under suitable conditions. The proof is made available upon request.

its welfare impact. As shown, top-revelation increases welfare beyond simple certification. It alleviates the Spencian distortion by pooling lower qualities while preserving informational gains and variety at the top of the distribution. The underlying design is almost identical to the optimal coarse certification but has a distinct distributional impact. Under top-revelation, fewer contracts are certified, and fewer consumers buy certified products. Revelation reduces insurers' gains from certification as consumers can distinguish between medium and high-quality certified products. Consumers, however, benefit from more information about certified products. Consumers in high-cost markets buy more uncertified products of lower quality, while those in low-cost and more competitive markets buy more certified products and thus benefit from the redesign.

Given the performance of top-revelation in the context of private information, it seems important to understand whether they can improve upon the optimal coarse monotone partitional design found in the main analysis. Appendix Figure 5 shows the optimal top-revealing design, and the last row of Panel A in Table 3 shows the resulting outcomes.[54] The design uses only four scores and has an aggregator imperfectly aligned with consumers' preferences. Like the coarse design, the bottom score pools quality at the bottom and has virtually no demand. Also, the aggregator is better aligned with consumers' preferences than the baseline to lower the losses from misclassification and multitasking moral hazard. However, unlike the optimal coarse design, consumers and contracts are more evenly spread across the second, third, and fully-revealing fourth scores. As heterogeneous firms locate themselves at different scoring thresholds, cost heterogeneity becomes more relevant for the design, leading to different optimal quality aggregators. This new design increases the welfare gains of redesigning the system by 4.5% but at an unknown complexity cost. This result confirms the value of the theoretical work on scores while simultaneously quantifying the moderate losses of adapting simpler approximations to the theoretical optimum.

VII.E Alternative regulatory objectives

Panel C of Table 3 and Appendix Figure 6 show the optimal monotone partitional designs under alternative regulatory objectives. They share key features with the main result: They all improve quality, information, and welfare relative to the status quo; They all pool quality at the bottom and use fewer scores than the baseline design; And they all improve the alignment of the aggregator weights with consumer preferences, albeit at different degrees. The aggregators are the key distinctive feature of each design, highlighting aggregation's

[54]This top-revealing design was optimized subject to the constraint of at most 9 scores and a linear aggregator.

critical role in shaping welfare outcomes. For example, the consumer surplus optimal design ($\rho^G = \rho^F = 0$) shifts weight from Access into Outcomes, as consumers value it more, and it is cheaper to produce at lower levels. This design matches more consumers at lower quality levels at substantially lower prices while minimizing the loss in coverage benefits. Coincidentally, it is also the design that expands the market the least. It induces balanced improvements across plans, leading to fewer changes large enough to offset TM consumers' systematic preferences for the public option. However, the benefits of this design are more evenly spread across MA consumers, improving consumer surplus far beyond any other design. Overall, the results show that regardless of the true regulatory objective, the key insights of the analysis stand, and substantial improvements could be attained.

VII.F Regulatory preferences

CMS's preferences over equilibrium quality might include factors beyond consumers' surplus and firms' profits. For example, they might believe that consumers undervalue the impact of letting their chronic conditions deteriorate because CMS is the residual payor for the associated expenses. This would help explain why the largest discrepancy between the optimal design and the baseline is the relative weight placed on chronic condition management (Intermediate) relative to medical quality (Outcome). As weights affect quality provision, CMS might be skewing the weight to shift the market towards their preferred outcome.

While the value of shifting the market is known only to CMS, the cost of doing so can be estimated. To do so, I compute the optimal certification design for a range of weights, starting from the optimum and adjusting the relative importance of the Intermediate and Outcome categories to span CMS's designs between 2009 and 2019. The results, shown in Appendix Figure 7, indicate that increasing the contribution of Intermediate relative to Outcome leads to a reallocation of investments from the second category to the first. However, the relative quality improvement grows slowly while consumers' WTP for certified products rapidly deteriorates. Certification becomes less representative of the information consumers need, and its effect on enrollment decreases. The drop in demand for certified products erodes investment incentives and quality plummets. To justify the design distortion for 2015, CMS would have to value a small improvement in chronic condition management 12 times more than what it costs to produce. It is outperformed by any subsidy that generates more than eight cents in investment per dollar spent.[55] Scores are a poor nudging mechanism

[55]The welfare loss is \$1.9 billion, computed by multiplying the relative loss from the 2015 design weights (23.5%) with the gains of the optimal design (\$155.74) and again by the number of Medicare beneficiaries in 2015 (54 million). The average investment in Intermediate increases by 0.21 million per contract.

as it is inherently costly to steer consumers with information they do not value.

VII.G Optimal design under the ignorance assumption

The results thus far have relied on the assumption that consumers understand changes in the scoring design. To understand this assumption's role in the results, I find the optimal design subject to the *ignorance* assumption. In this problem, the regulator knows consumers' quality preferences (e.g., through external surveys) but cannot change their perspective on what the star ratings mean. Instead, consumers' WTP for ratings is fixed at its estimated baseline value. Panel D of Table 3 and Appendix Figure 8 show the results.

The regulator's problem, in this case, can be thought of as labeling products using the Star Rating system, assigning each the "reputation" (i.e., subjective expected quality) associated with a baseline rating. The only constraint on this design is that the labeling must be weakly monotonic in quality. This labeling technology vastly improves the power of pooling qualities at the bottom, as the regulator can now assign all low qualities to the reputation of a baseline 1-star plan regardless of the breadth of the bottom scoring interval. The main downside of eliminating the informational channel is that the regulator cannot offer high-quality plans a reward greater than the perception of a baseline five-star plan. Therefore, relative to a regulator facing sophisticated consumers, one facing naive consumers can improve quality at the bottom of the distribution more and qualities at the top of the distribution less.

In total, welfare is greater when consumers are naive because the regulator's ability to coordinate demand to offset supply-side distortions is improved. The loss in information value is mitigated because the regulator is concerned with consumers' true valuation for quality and not their naive WTP for ratings. By labeling products correctly, the regulator can ensure consumers with higher WTP for quality are matched with higher-quality products. However, it must be noted that the evidence from Section IV rejects the assumption of ignorance. Thus, this exercise provides a bound on the gains from redesigning the system rather than an alternative plausible design. If consumers are partially informed, it stands to reason that the gains from redesign lie between these results and those of the main analysis.

An important caveat to this analysis is that the informed choice assumption was used to identify consumers' quality preferences. Without external data and under the assumption of ignorance, the regulator can only infer a lower bound on consumers' preferences ($\underline{\gamma}$). In Appendix III.G, I show that the methodology developed here can also be used to solve the robust scoring design problem $\max_{\psi\in\Psi}\min_{\gamma\geq\underline{\gamma}} TW(\psi,\rho^F,\rho^G|\gamma)$. The results show that the worst-case total welfare of the baseline design is slightly better than the best linear monotone

partitional score. Therefore, CMS's non-linear design offers a higher lower bound on welfare in the presence of preference uncertainty. This observation suggests that CMS's design might be driven by an abundance of caution about misrepresenting consumers' preferences. Given Medicare's delicate social and political role, this finding appears reasonable. However, this welfare advantage of the baseline design is attained under the assumption that consumers are entirely ignorant of changes to the scoring policy, which is rejected by the data.

Overall, these analyses complement the previous sections in three ways. First, they disentangle the mechanisms by which scores affect the market. In particular, designs in the main analysis coordinated consumers by changing the assignment of scores to products and consumers' beliefs about the quality represented by scores. This exercise eliminates the second channel, showing that scores can be effective even if consumers are unaware of design changes. Second, it provides a rationale for how the baseline design might have been formulated. This relies on assumptions rejected by the data and a severe aversion to misrepresenting consumers' preferences. Therefore, the welfare value of the proposed alternatives stands. Finally, it provides an alternative solution for the cautious regulator (or reader) unnerved by the assumption of informed choice.

VII.H Discussion

The results above have implications for scoring design beyond MA. The finding that optimal granularity is second order to optimal weighting indicates that the most salient feature of scores might be the least relevant one. This suggests that optimizing certifications might be better than disclosing more granular information in markets with moderate heterogeneity in WTP and good outside options (e.g., incremental technologies). This finding also implies a contradiction between efforts to regulate and disclose quality as neither quality nor consumers' information about it is monotonic in the ex-ante informativeness of scores.[56] More granular systems can lead to worse quality outcomes and exacerbate the effect of investment risk on quality variance. This is relevant for the joint efforts of CMS to promote and disclose quality (MedPAC, 2018) and likely also for many other markets where pay-for-performance and scoring policies coexist, such as in schooling, hospitals, and energy-efficient construction.

Finally, the empirical scoring design methodology developed in this article provides a solution to the gaming incentives that have plagued various disclosure policies (Feng Lu, 2012; Reynaert and Sallee, 2021). The results show that firms' incentives can be aligned with regulatory objectives by designing aggregation weights properly. Moreover, they can

[56]Here, ex-ante means when the designer chooses its scoring policy.

allow heterogeneous firms to reach high-quality production through various paths, which enables an array of products that stricter minimum quality standard policies would not permit. However, demand penalties must be imposed on those falling below a threshold to induce meaningful total investment. This finding contradicts recent advice given to Congress regarding eliminating "cliff effects" in insurer incentives in MA (MedPAC, 2020).

VIII Conclusion

This article studies the problem of designing a scoring system in the presence of market power. Using detailed data from Medicare Advantage in 2009-2015, I show that scores shift demand across products and alter insurers' investments. Exploiting variation in the scoring design, I solve the problem of a welfare-maximizing regulator, finding a constrained optimum and deriving findings about the scoring design problem by decomposing the solution.

The results suggest that optimal designs involve coarsening consumers' information. Under full information, market power over quality leads firms to invest inefficiently. A coarse score corrects these incentives by shifting demand, creating penalties for underperforming firms. This can be accomplished by simple and easy-to-interpret designs, such as binary certifications. Hence, there is no inherent conflict between scoring for sophisticated or more naive consumers; they both react to scores, change their demands, and exert regulatory pressure on firms. The results also show that using scores to steer quality production away from consumers' preferences can be extremely costly. Skewing scores' informational content quickly erodes their informational value and regulatory power. Finally, both theoretical and empirical results show that transparency in scoring design is paramount for eliciting consumers' preferences and the score's effectiveness as an informational policy.

My results support the growing theory on scoring design and point the way to several potential extensions. Incorporating market dynamics and measurement error would be helpful for scoring design in several markets with persistent investments and hard-to-measure outcomes. Accounting for data manipulation would help address challenges documented in nursing home scores and credit ratings. Finally, I assume that the quality domains and dimensions are fixed. How to define quality as a policy decision remains an open question.

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.

— and Gruber, J. (2011). Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program. *American Economic Review*, **101** (4), 1180–1210.

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.

Alé-Chilet, J. and Moshary, S. (2022). Beyond Consumer Switching: Supply Responses to Food Packaging and Advertising Regulations. *Marketing Science*, **41** (2), 243–270.

Allcott, H. (2011). Consumers perceptions and misperceptions of energy costs. *American Economic Review*, **101**, 98104.

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

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.

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

Atal, J. P., Cuesta, J. I. and Sæthre, M. (2022). Quality regulation and competition: Evidence from pharmaceutical markets. *Working paper*.

Aumann, R. J., Maschler, M. and Stearns, R. E. (1995). *Repeated games with incomplete information*. MIT press.

Ball, I. (2020). Scoring Strategic Agents. *Working paper*.

Barahona, N., Otero, C. and Otero, S. (2022). Equilibrium Eects of Food Labeling Policies. *Working paper*.

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

—, 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.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, pp. 265–272.

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

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.

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

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.

CMS (2016). Quality Strategy. *Technical report*.

Cooper, Z., Gibbons, S., Jones, S. and Mcguire, A. (2011). Does hospital competition save lives? Evidence from the English NHS patient choice reforms. *Economic Journal*, **121** (554), 228–260.

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

— 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.

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

Cutler, D. M., Huckman, R. S. and Kolstad, J. T. (2010). Input constraints and the efficiency of entry: Lessons from cardiac surgery. *American Economic Journal: Economic Policy*, **2** (1), 51–76.

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.

Dai, W. D., Jin, G., Lee, J. and Luca, M. (2018). Aggregation of consumer ratings: an application to Yelp.com. *Quantitative Marketing and Economics*, **16** (3), 289–339.

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.

Decarolis, F., Guglielmo, A. and Luscombe, C. (2020a). Open enrollment periods and plan choices. *Health Economics*, **29** (7), 733–747.

—, Polyakova, M. and Ryan, S. P. (2020b). Subsidy design in privately provided social insurance: Lessons from medicare part d. *Journal of Political Economy*, **128** (5), 1712–1752.

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

— 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.

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

Fioretti, M. and Wang, H. (2021). Performance Pay in Insurance Markets: Evidence from Medicare. *The Review of Economics and Statistics*, pp. 1–45.

Fleitas, S. (2020). Who benets when inertia is reduced? Competition, quality and returns to skill in health care markets. *Working paper*, p. 60.

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

Gaynor, M., Moreno-Serra, R. and Propper, C. (2013). Death by market power: Reform, competition, and patient outcomes in the national health service. *American Economic Journal: Economic Policy*, **5** (4), 134–166.

Geruso, M. and Layton, T. (2020). Upcoding: Evidence from medicare on squishy risk adjustment. *Journal of Political Economy*, **128**, 9841026.

Glazer, J. and McGuire, T. G. (2006). Optimal quality reporting in markets for health plans. *Journal of Health Economics*, **25**, 295–310.

—, —, Cao, Z. and Zaslavsky, A. (2008). Using global ratings of health plans to improve the quality of health care. *Journal of Health Economics*, **27**, 1182–1195.

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., Hendel, I. and Whinston, M. D. (2015). Equilibria in Health Exchanges: Adverse Selection versus Reclassification Risk. *Econometrica*, **83** (4), 1261–1313.

Handel, B. R. and Kolstad, J. T. (2015). Health insurance for humans: Information frictions, plan choice, and consumer welfare. *American Economic Review*, **105** (8), 24492500.

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

Ho, K. and Handel, B. (2021). Industrial organization of health care markets. *NBER Working Paper*.

— and Lee, R. S. (2017). Insurer Competition in Health Care Markets. *Econometrica*, **85** (2), 379–417.

Holmstrom, B. (1982). *Moral Hazard in Teams*. Tech. Rep. 2.

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

Hopenhayn, H. and Saeedi, M. (2019). Optimal Ratings and Market Outcomes. *NBER Working Paper Series*, pp. 1–39.

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

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.

— and Sorensen, A. T. (2006a). Information and consumer choice: The value of publicized health plan ratings. *Journal of Health Economics*, **25** (2), 248–275.

— and — (2006b). Information and consumer choice: The value of publicized health plan ratings. *Journal of Health Economics*, **25**, 248275.

Kamenica, E. (2019). Bayesian Persuasion and Information Design. *Annual Review of Economics*, **11** (1), 249–272.

— 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*.

Kolotilin, A. and Zapechelnyuk, A. (2019). Persuasion meets delegation. *arXiv preprint arXiv:1902.02628*.

Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, **103** (7), 2875–2910.

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. (2009). Measuring welfare losses from adverse selection and imperfect competition in privatized medicare. *Working paper*.

Malherbe, C. and Vayatis, N. (2017). Global optimization of Lipschitz functions. *34th International Conference on Machine Learning, ICML 2017*, **5** (1972), 3592–3601.

Marone, V. R. and Sabety, A. (2022). When should there be vertical choice in health insurance markets? *American Economic Review*, **112** (1), 304–42.

McGuire, T. G. and Newhouse, J. P. (2018). *Chapter 19 - Medicare Advantage: Regulated Competition in the Shadow of a Public Option*, Academic Press, p. 563598.

—, — and Sinaiko, A. D. (2011). An economic history of Medicare Part C. *Milbank Quarterly*, **89** (2), 289–332.

MedPAC (2018). The Medicare Advantage Program: Status Report. *Report to Congress: Medicare Payment Policy*, pp. 287–306.

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

Miller, K. S., Petrin, A., Town, R. and Chernew, M. (2022). Optimal managed competition subsidies. *NBER Working Paper Series*.

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

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

—, Price, M., McWilliams, J. M., Hsu, J. and McGuire, T. G. (2015). How much favorable selection is left in medicare advantage? **1**, 126.

Nosal, K. (2011). Estimating Switching Costs for Medicare Advantage Plans. *Working paper*, pp. 1–45.

Population Health Institute, U. o. W. (2024). County health rankings & roadmaps 2024. https://www.countyhealthrankings.org/.

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.

Reimers, I. and Waldfogel, J. (2021). Digitization and pre-purchase information: The causal and welfare impacts of reviews and crowd ratings. *American Economic Review*, **111**, 19441971.

Reynaert, M. and Sallee, J. M. (2021). Who benefits when firms game corrective policies? *American Economic Journal: Economic Policy*, **13** (1), 372–412.

Ryan, C. (2020). How does Insurance Competition Affect Medical Consumption? *Working paper*.

Schennach, S. M. (2016). Recent Advances in the Measurement Error Literature. **8**, 341–377.

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**, 105130.

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

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.

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