

## QUALITY DISCLOSURE AND REGULATION: SCORING DESIGN IN MEDICARE ADVANTAGE

BENJAMIN VATTER

Sloan School of Management, Massachusetts Institute of Technology

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.

### 1. INTRODUCTION

QUALITY SCORES ARE UBIQUITOUS. From healthcare to food quality, car emissions, or school performance, regulators and certifying agencies rely on scores for disclosure. Scores help consumers choose when information is scarce or difficult to process 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 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 harm consumers (see, e.g., [Silver-Greenberg and Gebeloff \(2021\)](#)). 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.

---

Benjamin Vatter: [bvatter@mit.edu](mailto:bvatter@mit.edu)

I thank David Dranove, Igal Hendel, Gaston Illanes, and Amanda Starc for their invaluable mentorship and advice. I thank Joseph Doyle, 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. Funding Statement: Open Access funding enabled and organized by MIT Hybrid 2025.

© 2025 The Author. *Econometrica* published by John Wiley & Sons Ltd on behalf of The Econometric Society. Benjamin Vatter is the corresponding author on this paper. This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](#), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

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. The MA setting provides a valuable laboratory for studying disclosure design: The rules mapping quality measurements to scores—that is, 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, Bravo, Hull, and Starc 2021) and entails billions in public spending (CMS (2016)).

I document three fundamental observations about scoring design in MA. First, consumers have increasing preferences for scores: New enrollees are 18% 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 identify the primitives of an empirical model of quality investment, plan pricing, and enrollment. Consumers’ willingness to pay (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 changes insurers’ relative gains from investing in different quality dimensions, identifying their investment costs.

The model captures 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, Shcherbakov, and Shum (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 WTP for quality to match with diverse products, but also increases the potential for inefficient quality production.

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.<sup>1</sup> 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 equilibrium responses are accounted for.

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 gains under the new scores are 17% larger than under full information.

This exercise in empirical scoring design bridges a gap between the theoretical literature on the subject and the empirical literature that measures disclosures' impact.<sup>2</sup> To my knowledge, few papers have explored this gap. Dai, Jin, Lee, and Luca (2018) studied the optimal aggregation of subjective consumer restaurant reviews and Blattner, Hartwig, and Nelson (2022) the design of credit scores. Closely related, Barahona, Otero, and Otero (2023) examined the equilibrium effect of food labeling and the design of the threshold

<sup>1</sup>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.

<sup>2</sup>Theoretical work includes, among others, Albano and Lizzeri (2001), Glazer and McGuire (2006), Harbaugh and Rasmusen (2018), Hopenhayn and Saeedi (2019), Ball (2025); and Zapechelnyuk (2020). The empirical work includes, among others, Jin and Sorensen (2006), Elfenbein, Fisman, and McManus (2015), Araya, Elberg, Noton, and Schwartz (2018), Alé-Chilet and Moshary (2022) and Reynaert and Sallee (2021). See Dranove and Jin (2010) for a review of earlier work on disclosure and Kamenica (2019) for work on theoretical information design.

for breakfast cereal to be certified as unhealthy. As in this article, they found that supply-side responses augment the effectiveness of the labels and can outperform more informative policies. This article 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. In addition, I provide a novel identification result showing that variation in scoring design can be used to identify consumers' beliefs about quality, which can serve to complement or substitute the survey-based elicited-beliefs approach of Barahona, Otero, and Otero (2023).

The 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, Gallego, and Neilson (2019)) and the empirical study of quality regulation (Angrist and Guryan (2008), Kleiner and Soltas (2019), Larsen, Ju, Kapor, and Yu (2020), Atal, Cuesta, and Sæthre (2022)). My examination of the regulation of imperfect competition among insurers expands on the literature on quality provision in healthcare markets (Cutler, Huckman, and Kolstad (2010), Cooper, Gibbons, Jones, and McGuire (2011), Gaynor, Moreno-Serra, and Propper (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.

As the MA scores are designed to simplify enrollment decisions, this work relates to a large literature on choice frictions in health insurance. This literature has documented that consumers often choose dominated plans (Abaluck and Gruber (2011)), fail to understand cost-sharing rules (Handel and Kolstad (2015)), or are inertial and inattentive (Handel (2013), Polyakova (2016), Ho, Hogan, and Morton (2017)). This work complements research on how lack of information about complex plan features results in suboptimal enrollment and how regulation might alleviate these frictions. In particular, in both Brown and Jeon (2024) and this article, consumers make enrollment mistakes due to limited information, and regulation can alleviate these errors (in part) by eliminating suboptimal choices. In Brown and Jeon (2024), suboptimal plans are eliminated to reduce the informational burden on consumers, while in this article, they are eliminated through low scores to incentivize efficient quality investment.

Finally, this paper connects research on the industrial organization of MA (Town and Liu (2003), Aizawa and Kim (2018), Curto, Einav, Levin, and Bhattacharya (2021), Ryan (2020), Decarolis, Guglielmo, and Luscombe (2020), Miller, Petrin, Town, and Chernew (2022)) to the literature on insurance market design (Handel, Hendel, and Whinston (2015), Decarolis, Polyakova, and Ryan (2020), Marone and Sabety (2022)). I study the role of 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) studied 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.

## 2. 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. Consider a single-product monopolist selling an indivisible good characterized by two dimensions of quality,  $q = (q_1, q_2)$ , and a price. The monopolist chooses quality, paying an increasing

and convex investment cost, and then selects a price for the product. A regulator observes the product's quality and discloses a public score (or signal)  $\psi(q)$  to the market. Consumers cannot observe the product's quality but know the regulator's scoring rule and the realized score. Using this information and knowing that quality is costly to produce, consumers form rational expectations about the vector  $q$  and make purchasing decisions. The regulator seeks to maximize welfare by committing to a policy before the monopolist's quality is chosen.

The regulator can attain two informational extrema. On the one hand, a constant score reveals no information to consumers, rendering demand inelastic to quality, thus eliminating incentives for the monopolist to invest. On the other hand, a fully informative score allows the monopolist to exert market power over quality (Crawford, Shcherbakov, and Shum (2019)), leading to potentially inefficient investment (Spence (1975)). Intuitively, when evaluating a marginal quality increase under full information, the monopolist considers its effect only on the marginal consumer. Efficiency, instead, requires accounting also for the surplus created for inframarginal consumers. Hence, the monopolist's investment will likely be inefficient, even under full information.<sup>3</sup> Figure 1(a) illustrates these two extrema and the resulting inefficiencies as a function of the choice of  $q_1$ . Figure 1(b) shows the distortion in quality space, revealing that the *Spencian* distortion is operating on both quality dimensions.

The regulator can address these inefficiencies by using coarse scores. Figure 1(c) illustrates the outcome of a scoring rule certifying whether a weighted average of quality ( $\psi(q) = \omega_1 q_1 + \omega_2 q_2$ ) exceeds a threshold ( $\bar{\psi}$ ). As shown in Figure 1(d), the threshold and weights are chosen such that the boundary between scores is tangent to consumers' iso-utility curve at the full-information welfare-optimal quality,  $q^w$ . This policy disrupts the firm's profit curve because, on both sides of the certification cutoff, demand has different levels but is inelastic to quality. 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, they expect  $q = 0$ . To the right, consumers are guaranteed that quality is at least on the scoring boundary. As the firm's isocost is tangent to the scoring threshold at  $q^w$ , the efficient outcome is also the most cost-effective investment for the firm to attain certification. Thus, consumers expect a certified product to have quality  $q^w$ . Therefore, if the monopolist's full-information profits at zero are lower than at  $q^w$ , it will invest efficiently when regulated. Consumers' expectations would then be accurate, thus eliminating market power over quality and informational distortions.

Thus, the optimal scoring policy resolves the monopolist's moral hazard problem by establishing a contract by which an efficient investment is rewarded with high demand and a suboptimal investment is penalized with low demand. Figure 1(d), however, reveals that the multidimensional nature of quality adds two additional complexities to the regulatory problem. First, since consumers cannot observe the combination of quality by which a product attained its score, firms' decisions ignore their preferences over quality. In the figure, the monopolist invests at  $q^w$  only because it is the cheapest way to attain certification. If a firm with a different cost structure were to enter the market, its chosen quality might differ substantially from the optimal one. Thus, in a scored environment, firms only consider their costs when choosing an investment mix, while the regulator also values consumers' preferences, introducing a multitask moral hazard problem (Holmstrom and Milgrom (1991)).

<sup>3</sup>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.

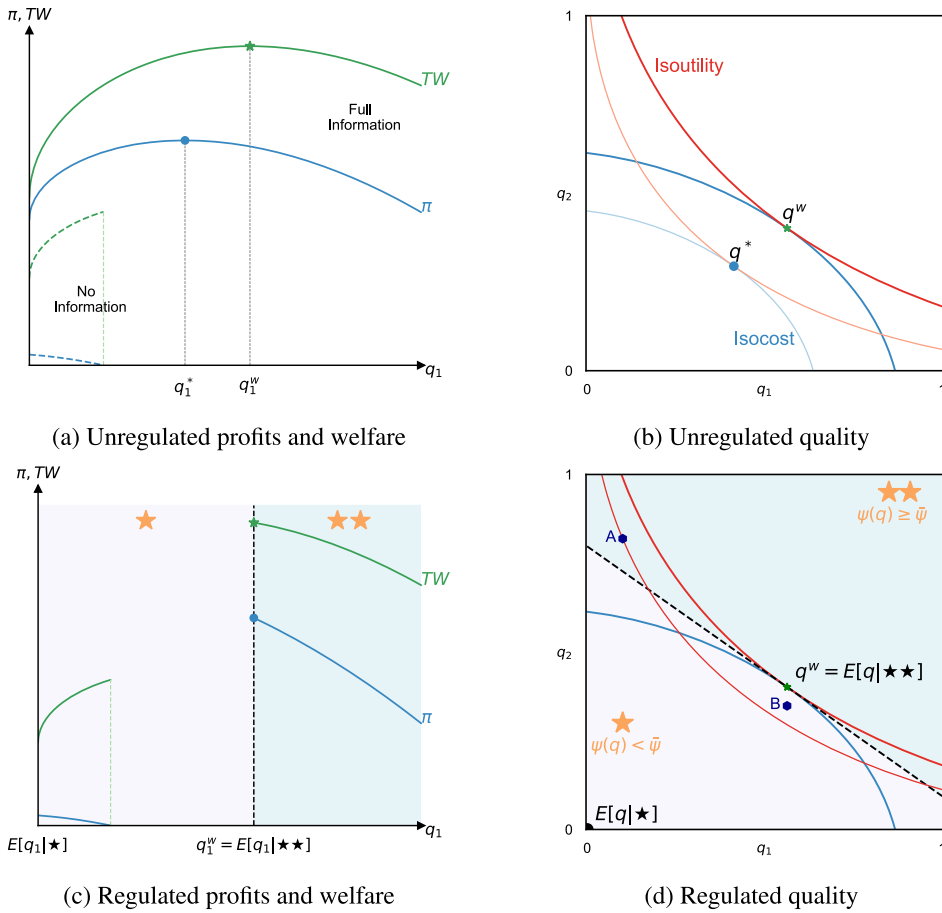


FIGURE 1.—Quality certification under monopolistic provision. *Note:* These figures illustrate how certification changes a monopolist’s investment incentives. Figure (a) presents profit and total welfare curves as a function of the first dimension of quality ( $q_1$ ), holding the second dimension fixed at its efficient level. Curves are shown for the full- and no-information scenarios, the latter vanishing when profits are negative as the monopolist exits. Figure (b) presents the same scenario in quality space, showing how consumers’ iso-utility curves and the firm’s indifference curves meet at the monopolist optimal investment ( $q^*$ ) and at the social optimum ( $q^w$ ). Figure (c) presents how profit and welfare change when consumers are only informed whether quality exceeds a threshold  $\psi(q) \geq \bar{\psi}$ . The monopolist’s profit curve in (c) is disrupted as 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. Figure (d) shows the same scoring threshold in quality space. Hexagons  $A$  and  $B$  mark two potential misclassified products under the optimal design:  $A$  would get certified, but consumers would prefer the uncertified  $B$ . The welfare optimum in all figures is the same. Shaded areas illustrate the distinct scores.

Second, unless scores aggregate quality precisely according to consumers’ preferences, consumers cannot tell whether a higher-scoring product is preferred to a lower-scoring one. For example, in Figure 1(d), a product of quality  $B$  would be preferred to one of  $A$ . The scoring system, however, would award the certification to  $A$  but not  $B$ . This potential for misclassification becomes particularly relevant when firms’ ability to control their quality becomes imperfect, as will be the case in the empirical application of this article.



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. As illustrated in the figures above, determining the optimal design entails recovering firms' investment costs and consumers' quality preferences, in addition to the usual demand and cost estimates governing prices and quantity. In the following sections, I develop a methodology to recover these components and systematically translate them into optimal scoring designs.

### 3. INSTITUTIONAL DETAILS AND DATA

#### 3.1. *Medicare Advantage and the Star Rating Program*

Since 1965, retirees and disabled individuals in the United States 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.<sup>4</sup>

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.<sup>5</sup>

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. 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

<sup>4</sup>TM consists of Part A (hospital coverage) and Part B (physician and outpatient coverage). For further details on the history of this program, see McGuire, Newhouse, and Sinaiko (2011).

<sup>5</sup>MA consumers pay a Part B premium regardless of their choice of TM or MA. For further details regarding the MA market regulation, see Supplemental Appendix 1.1 (Vatter (2025)).

each plan's quality next to the enrollment button in Medicare's unified shopping platform.<sup>6</sup> 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

$$\text{Score}_j = \text{Round}_{0.5} \left( \frac{\sum_{k \in \mathcal{K}} w_k \sum_{l \in \mathcal{L}_k} \text{MeasureScore}_l(q_{lj})}{\sum_{k \in \mathcal{K}} w_k |\mathcal{L}_k|} + \omega_j \right), \quad (1)$$

where  $\text{Round}_{0.5}(\cdot)$  rounds a number to its nearest half and  $q_{kj}$  is the quality of plan  $j$  in measure  $k$ . The adjustment factor,  $\omega_j$ , captures minor bonuses due to past performance.<sup>7</sup>

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.<sup>8</sup> I call the total weight assigned to a category its *design contribution*. These contributions have varied significantly, as shown in Figure 2(a). As I detail in Supplemental Appendix 1.6, consumers likely observed this variation since the composition of categories was visible on the Medicare website and enrollment platform.<sup>9</sup>

Design variation significantly impacted score assignment. Figure 2(b) shows that if CMS had kept its 2011 scoring design, 60% of plans in 2019 would have received 4 or more 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.<sup>10</sup> However, this paper aims to

<sup>6</sup>See Supplemental Appendix 1.2 for a description of the online platform (Vatter (2025)). For a description of earlier quality scores in MA, see Dafny and Dranove (2008).

<sup>7</sup>See the Supplemental Material (Vatter (2025)) 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 Supplemental Appendix 2.10 for evidence against the influence of quality selection and manipulation (Vatter (2025)).

<sup>8</sup>See the Supplemental Material for correlation within and across categories.

<sup>9</sup>Throughout, the Supplemental Appendix refers to the online supplementary material (Vatter (2025)).

<sup>10</sup>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.



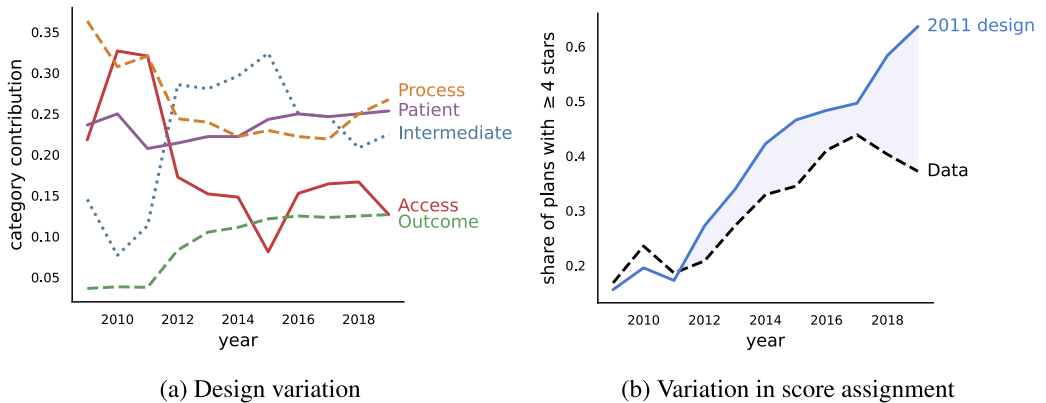


FIGURE 2.—Scoring design variation and simulated assignment under constant design. *Note:* Figure (a) shows the evolution of the scoring design category contributions. Each is the product of the category's number of measurements (e.g., Process includes breast cancer screening and kidney disease monitoring) 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.

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.

### 3.2. 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.<sup>11</sup>

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.<sup>12</sup> The data include 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.<sup>13</sup> I restrict the data to the continental United

<sup>11</sup>Similar restrictions have been used by Aizawa and Kim (2018), Curto et al. (2021), and Miller et al. (2022). I present descriptive statistics in the Supplemental Appendix Table 1.

<sup>12</sup>Excluding 2014, because it was never released to the public due to implementation difficulties.

<sup>13</sup>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 Supplemental Material.

States, 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 come from Medical Loss Ratio filings made by MA insurers, which recent regulation changes have modified to include a separate item for quality investment.<sup>14</sup>

The fifth and final data source is the University of Wisconsin's County Health Rankings (Population Health Institute, University of Wisconsin (2024)), which contains vital information about the availability 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.

I use these sources to estimate the primitives governing the scoring design problem, including consumers' preferences and beliefs about quality and insurers' investment costs. The next section provides evidence of the effects of scoring design, showing consumers respond to scores, have some understanding of the scoring design, and that insurers respond to scoring incentives. It documents the variation used to identify the empirical model used to solve the scoring design problem and provides support for some of its key assumptions.

#### 4. EVIDENCE ON MARKET RESPONSES TO SCORING

##### 4.1. *Scoring and Enrollment*

Beneficiaries' responsiveness to scores has been thoroughly documented in previous work (Dafny and Dranove (2008), Reid, Deb, Howell, and Shrank (2013), Darden and McCarthy (2015)) and is noticeable in individuals' enrollment decisions.<sup>15</sup> To document this effect, I regress an indicator of each new potential MA enrollee's choice on the plan's score and plan attributes:

$$y_{ijt} = \alpha_{r(jt)} + \mathbf{x}_{jt}\boldsymbol{\lambda} + \mu_{it} + \epsilon_{ijt}. \quad (2)$$

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  $t$ , and  $\mathbf{x}_{jt}$  is a vector of plan characteristics including premium, benefit levels, Part D coverage, and indicators for additional vision, hearing, and dental benefits. In addition,  $\mathbf{x}_{jt}$  includes a fixed effect for TM to capture its overall benefit level and average quality. The choice-event (beneficiary-year) fixed effect,  $\mu_{it}$ , normalizes the effect of plan attributes and scoring relative to the options available to each consumer.

The first column of Table I shows the estimates of  $\alpha_{r(jt)}$ . All else equal, consumers prefer higher-scoring plans: A new enrollee is approximately 4.5 percent more likely to enroll

<sup>14</sup>MLR regulation is not binding in MA and hence ignored during the analysis (Curto, Einav, Finkelstein, Levin, and Bhattacharya (2019)).

<sup>15</sup>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.

TABLE I  
ENROLLMENT RESPONSES TO SCORING DESIGN.

		(1)	(2)	(3)
<u>Star Rating</u>	3	0.006 (0.001)		
	3.5	0.012 (0.001)		
	4	0.021 (0.001)		
	4.5	0.020 (0.001)		
	5	0.045 (0.002)		
<u>High Rated (<math>\geq 4</math>)</u>	Baseline		−0.009 (0.003)	−0.008 (0.003)
	$\hookrightarrow \times$ Chronically ill		0.014 (0.003)	0.013 (0.003)
	$\hookrightarrow \times$ Diabetic		0.019 (0.003)	0.018 (0.003)
<u><math>\hookrightarrow \times</math> Measure Weights</u>	Baseline		0.012 (0.010)	0.011 (0.017)
	$\hookrightarrow \times$ Chronically ill		0.032 (0.011)	0.061 (0.019)
	$\hookrightarrow \times$ Diabetic		0.050 (0.013)	0.089 (0.021)
<i>N</i>		416,399	416,399	416,399
<i>R</i> <sup>2</sup>		0.747	0.747	0.747
Measures			Intermediate	Diabetic Only

*Note:* This table shows estimates from Equations (2) and (3). The sample includes all potential new MA enrollees, and the unit of observation is a member-plan option, including TM. Regressions include controls for plan premium, benefits, Part D coverage, indicators for additional vision, dental, and hearing benefits, a TM fixed effect, and a choice-event (member-year) fixed effect. The first column presents the effect of plan star ratings on enrollment relative to plans with a rating less than or equal to 2.5 stars. The second and third columns present interactions between an indicator of high rating ( $\geq 4$  stars), the weight given in the current design to a subset of measures, and indicators of chronic illness. In column (2), the subset is all Intermediate Outcome category measures. Column (3) is all measures related to diabetic care. For a full list of chronic conditions documented in the MCBS data, see Supplemental Appendix 1.5. New plans without ratings are excluded. Regressions weighted by MCBS sampling weights. Homoscedastic standard errors are in parentheses.

in a 5-star plan than in an equivalent plan of 2.5 stars or less (the normalized category). Conditional on enrolling in any MA plan, this incremental effect corresponds to an 18 percent increase in enrollment probability (see Supplemental Appendix Table 2). The effect of scores is monotonic, as expected if consumers understand that scores signal quality.

#### 4.2. Consumers' Understanding of Scoring

Despite well-documented responses to scores in enrollment, whether consumers are aware of the scores' design and can interpret them correctly is unknown. Possibly, enrollees have only an intuitive understanding that scores signal quality but are otherwise ignorant of the regulator's design. This hypothesis is testable. Under the null of ignorance, enrollment decisions should be unaffected by policy parameters whose sole effect is to change the interpretation of the star ratings. Moreover, if those policy parameters change such that higher ratings are given to plans that are beneficial to specific populations, policy-ignorant consumers should not be able to take advantage of the change. To operationalize these ideas, I examine whether consumers who are chronically ill are more likely to buy high-rated plans when more weight is given to quality measures related to chronic condition management. In particular, I estimate the regression

$$y_{ijt} = \sum_{l \in \mathcal{L}} (\alpha_l + \beta_l \omega_{kt}) \mathbb{1}\{l(i) = l\} \mathbb{1}\{r(jt) \geq 4\} + \mathbf{x}_{jt} \boldsymbol{\lambda} + \mu_{it} + \epsilon_{ijt}, \quad (3)$$

where groups  $\mathcal{L}$  and  $l(i)$  indicate whether the consumer has no reported chronic conditions, has any chronic condition except for diabetes, or is diabetic. The policy parameter

$\omega_{kt}$  captures the relative weight given to a subset of measures  $k$  in year  $t$ 's design. Thus,  $\alpha_l$  captures the group's propensity to enroll in highly-rated plans and  $\beta_l$  how this propensity changes as the scoring design changes. Under the null of policy ignorance,  $\beta_l$  should be zero.

The second column of Table I shows that chronically ill consumers are more likely to buy a high-rated plan when scores are more reflective of chronic condition management. The third column focuses on design changes that increase the importance of diabetic care in the design, showing that diabetic beneficiaries are particularly responsive to such changes. As noted in the table, non-diabetic chronically ill consumers also value diabetes management, likely due to their high likelihood of developing diabetes in the future.

These results reject the null of policy ignorance and support the hypothesis that consumers are aware of the regulator's scoring policy. Supplemental Appendix 1.6 provides additional supporting evidence, showing that the likelihood of enrolling in high-rated plans changed following large and well-publicized design changes in 2012 and correlates with changes to category contributions. The Appendix also shows that consumers are correct in valuing scores as beneficiaries enrolled in higher-scoring plans have better outcomes. Therefore, given the evidence, the main analysis conducted in this article assumes consumers are aware of changes to the scoring design and can interpret them correctly. I present results under the alternative hypothesis of policy ignorance in Supplemental Appendix 4. Importantly, even if consumers are unaware of policy changes, they have stated preferences for higher-rated products, and therefore, the regulator can steer demand toward high-quality plans and change firms' investment incentives. Thus, the main findings of this paper hold even under the assumption of policy ignorance.

### 4.3. *Supplied Quality Responses to Scoring*

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 3(a). 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.<sup>16</sup> Finally, because the scoring rule converts quality measurements to measure-level scores, the change produced clear and heterogeneous incentives across firms. For example, Figures 3(b) and 3(c) 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 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

<sup>16</sup>Changes were announced a year before measurement, allowing insurers to respond in time.

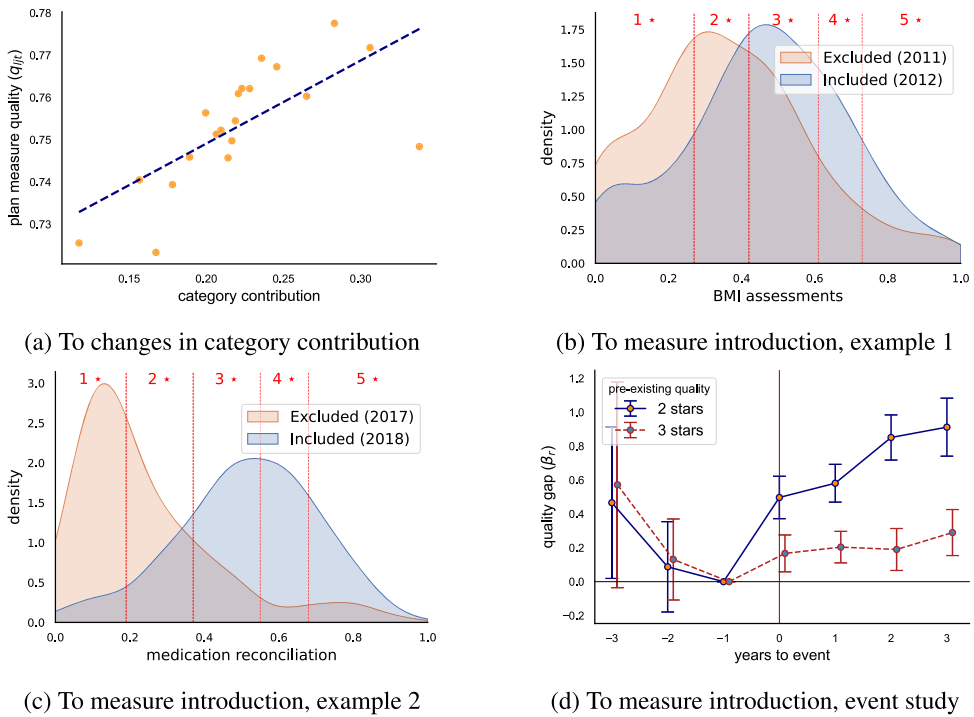


FIGURE 3.—Plan quality response to design variation. *Note:* Figure (a) is a binned scatter plot of quality at the contract measure year level and its correlation with each measure’s category contribution to that year’s design. Observations have been residualized against a plan-measure fixed effect to 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 (4), relative to the 4-star category and the year before introduction. Bars mark 95% confidence intervals.

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

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

Above,  $\tau$  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 ( $\tau = -1$ ).<sup>17</sup> To compare quality metrics, I standardize them using their means and standard deviations across all years. To avoid conflating the

<sup>17</sup>For example, in Figure 3(b), I classify a plan of measure quality 0.5 in the third group.

effects of bounded quality domains, I drop plans in the first and last preexisting quality groups and normalize the coefficient of interest ( $\beta_{rr}$ ) for the fourth group to zero.<sup>18</sup>

The analysis involves three differences. First is a comparison within plan-measure, controlled for by the fixed effect  $\gamma_{ij}$ . 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 ( $G_{ij}$ ) and the measure-time fixed effect  $\mu_{it}$ . In this case,  $\beta_{rr}$  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  $\xi_{jt}$ . 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 3(d) plots the  $\beta_{rr}$  estimates, and Supplemental Appendix 1.8 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.

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. The following section presents a model that rationalizes these scoring effects and allows me to leverage MA's extensive variation.

## 5. EMPIRICAL 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.

### 5.1. 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  $p_{jmt}^{\text{total}}$ , cost-sharing benefits level  $b_{jmt}$ , additional plan attributes  $\mathbf{a}_{jmt}$  (e.g., bundled dental

<sup>18</sup>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 Supplemental Appendix 1.8 for robustness.



insurance), and a score of  $r_{jt}$ . The expected indirect utility of consumer  $i$  from plan  $j$  is

$$u_{ijmt} = \underbrace{\alpha_i p_{jmt}^{\text{total}}}_{\text{premium}} + \underbrace{\beta_i b_{jmt}}_{\text{benefits}} + \underbrace{\mathcal{E}[v(\mathbf{q})|r_{jt}, \psi_t]}_{\text{quality}} + \underbrace{\lambda^{a'} a_{jmt}}_{\text{plan attributes}} + \underbrace{\lambda^l l_{ijt}}_{\text{demographic preferences}} + \underbrace{\xi_{jmt}}_{\text{unobserved preference}} + \underbrace{\varepsilon_{ijmt}}_{\sim \text{T1EV}}. \quad (5)$$

Consumers have heterogeneous preferences for premiums and benefits ( $\alpha_i$ ,  $\beta_i$ ). 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 the plan, such as copayments and coinsurance, and are shown on the enrollment platform. Total premiums include Part C (MA) and Part D (prescription-drug) premiums associated with the plan.<sup>19</sup> Consumers value the vector of quality  $\mathbf{q}$  at  $v(\mathbf{q})$ , forming subjective expectations about this vector given the scoring policy ( $\psi_t$ ) and the plan's score ( $r_{jt}$ ).<sup>20</sup> Consumers also value the plan's bundled services ( $\lambda^a$ ) and have systematic preferences for certain insurers and MA overall based on their demographic group ( $\lambda^l$ ). Finally, consumers have unobserved preferences for plans ( $\xi_{jmt}$ ) and independent type-1 extreme value idiosyncratic preferences ( $\varepsilon_{ijmt}$ ), 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  $p_{0mt}^D$  the price of the market's most popular Part D plan.<sup>21</sup> Consumers' heterogeneous preferences for TM are captured by their demographic relative preferences for MA. The outside option's indirect utility is thus  $u_{i0mt} = \alpha_i p_{0mt}^D + \beta_i b_0 + \varepsilon_{i0mt}$ .

<sup>19</sup>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.

<sup>20</sup>Some approximations are required to derive this indirect utility. As shown by [Tebaldi, Torgovitsky, and Yang \(2023\)](#), for risk-averse consumers to have indirect utilities that are linear in premiums, they must have constant absolute risk aversion. In particular, consider  $U_{ijmt} = -\frac{1}{A_i} \exp[-A_i(\text{Inc}_{it} - p_{jmt}^{\text{total}} - m_{jmt} + v(\mathbf{q}_j) + \bar{\xi}_{ijmt})]$ , where  $A_i$  is the risk-aversion parameter,  $\text{Inc}_{it}$  income,  $m_{jmt}$  spending, and  $\bar{\xi}_{ijmt}$  absorbs the utility from all other product attributes. Assuming quality, risk protection, and spending are independent and additively separable (see Supplemental Appendix 2.2 for details on modeled quality), results in an indirect utility  $u_{ijmt} = p_{jmt}^{\text{total}} + \bar{\xi}_{ijmt} - \frac{1}{A_i} \ln(\mathbb{E}_m[\exp(A_i m_{jmt})]) - \frac{1}{A_i} \ln(\mathcal{E}_q[\exp(-A_i v(\mathbf{q}_j))|r_{jt}, \psi_t])$ . The indirect utility of quality satisfies  $-\frac{1}{A_i} \ln(\mathcal{E}_q[\exp(-A_i v(\mathbf{q}_j))|r_{jt}, \psi_t]) = \mathcal{E}_q[v(\mathbf{q})|r_{jt}, \psi_t] + \vartheta_{ijt}$ , where  $\vartheta_{ijt} \in [-A_i \frac{\Delta_{r_{jt}, \psi_t, v(\mathbf{q})}^2}{8}, 0]$ . The upper limit on  $\vartheta$  follows from Jensen's inequality and the lower from Hoeffding's Lemma, with  $\Delta_{r_{jt}, \psi_t, v(\mathbf{q})}$  being the maximum difference in quality-utility given score  $r_{jt}$  and design  $\psi_t$ . This article takes  $\vartheta = 0$  to simplify and make progress on the empirical scoring design question. To assess the error introduced by this approximation, we can use the value of  $A_i$  from [Handel \(2013\)](#) and the estimated maximum difference in quality-utility for four-star plans (see Section 6.1.2), which results in a lower bound for  $\vartheta$  of approximately  $-\$3.9$ . For comparison, the estimated value of quality is in the order of the tens of thousands of dollars. This limited error is due to the high granularity of the MA scoring system. Less granular scores (e.g., certifications) might incur larger errors from assuming indirect utilities that are linear in expected quality. Finally, in order to obtain a term linear in spending and, therefore, of benefits ( $b_{jmt}$ ), we can appeal to the approximation of [Abaluck and Gruber \(2011\)](#), namely, assuming that  $m$  has a normal distribution and taking a first-order Taylor approximation to the resulting indirect utility component. It is worth noting that previous work on the effect of scores (e.g., [Barahona, Otero, and Otero \(2023\)](#)) did not have to make these approximations as they are not set in an insurance market and could assume risk-neutral preferences.

<sup>21</sup>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.

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  $\gamma_{it}$  for risk-adjustment purposes. I denote the risk-adjusted demand of plan  $j$  as  $\tilde{D}_{jmt} = \sum_{i \in \mathcal{I}_{mt}} \gamma_{it} s_{ijmt}$ .

## 5.2. Supply

### 5.2.1. Pricing

Each year  $t$ , at the third stage of the game, insurance firm  $f$  observes the vectors of realized qualities  $\mathbf{q}_t$  and scores  $\psi_t(\mathbf{q}_t) = \mathbf{r}_t$ . Given this information, the firm chooses prices to maximize its total profits given by<sup>22</sup>

$$V_{fmt}(\mathbf{q}_t, \psi) = \max_{p_{fmt}} \sum_{j \in \mathcal{J}_{fmt}} \underbrace{\tilde{D}_{jmt}(\mathbf{p}_{mt}, \mathbf{r}_t)}_{\text{risk-adjusted demand}} \underbrace{(p_{jmt} + R(p_{jmt}, \mathbf{z}_{jt}))}_{\text{marginal revenue}} - \underbrace{C(\mathbf{q}_{jt}, \mathbf{a}_{jmt}, \boldsymbol{\theta}^c)}_{\text{marginal cost}}. \quad (6)$$

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 ( $\mathbf{z}_{jt}$ ), 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 Supplemental Appendix 2.1. 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 ( $\mathbf{q}_{jt}$ ), additional attributes as included in the demand ( $\mathbf{a}_{jmt}$ ), 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.

### 5.2.2. Investment

In the game's second stage, each firm observes the regulator's scoring rule  $\psi_t$  and chooses an investment level  $x_{\text{ckt}}$  for each of its contracts  $c \in \mathcal{C}_{ft}$  and category of quality  $k$ .<sup>23</sup> 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:

$$\pi_{ft}(\psi_t) = \max_{x_{ft}} \sum_m \underbrace{\int \mathbb{E}_{mt}[V_{fmt}(\mathbf{q}_f, \mathbf{q}_{-f}, \psi_t)] dF(\mathbf{q}_f | \mathbf{x}_{ft})}_{\text{expected insurance profit}} - \underbrace{I_f(\mathbf{x}_{ft})}_{\text{investment cost}}. \quad (7)$$

<sup>22</sup>In MA, this price is called a "bid." I avoid this terminology to prevent confusion with auctions.

<sup>23</sup>Each contract is associated with a set of plans  $\mathcal{J}_{ct}$  such that  $\mathcal{J}_{ft} = \bigcup_{c \in \mathcal{C}_{ft}} \mathcal{J}_{ct}$ .

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  $F(q_f|x_{fi})$ . 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 ( $q_{-f}$ ). 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 characteristics, and their previous enrollment choices. The assumption is motivated by the secrecy of insurers' contractual arrangements and the lack of investment data.<sup>24</sup>

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.

### 5.3. 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  $\psi$  from within a class  $\Psi$ :

$$\begin{aligned} & \text{TW}(\psi, \rho^F, \rho^G) \\ &= \int \left[ \underbrace{\text{CS}(\psi, q)}_{\text{Consumer surplus}} + \underbrace{\rho^F \sum_f V_f(\psi, q) - I_f(x_f^*(\psi))}_{\text{Insurer profit}} - \underbrace{\rho^G G(\psi, q)}_{\text{Government spending}} \right] dF(q|x^*(\psi)), \end{aligned} \quad (8)$$

where  $x^*(\psi)$  denotes the equilibrium investment induced by the scoring policy,  $G(\cdot)$  the governmental subsidy spending on TM services and MA enrollment, and  $\text{CS}(\cdot)$  consumers' surplus from enrollment.<sup>25</sup> 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:<sup>26</sup>

$$\text{CS}(\psi, q)$$

<sup>24</sup>I present evidence of imperfect quality control in Supplemental Appendix 1.10. The assumption about firms' beliefs is similar to that of Sweeting (2009). This article's limited investment data only recently became available.

<sup>25</sup>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.

<sup>26</sup>This surplus standard is common in the literature on choice under uncertainty. Similar approaches have been used by Jin and Sorensen (2006), Allcott (2011), and more recently Reimers and Waldfogel (2021).

$$= \sum_{i \in \mathcal{I}} \frac{1}{|\alpha_i|} \left( \underbrace{\ln \left( \sum_{j \in \mathcal{J}_{mt} \cup \{0\}} \exp(\delta_{ijmt}) \right)}_{\text{ex ante surplus}} - \underbrace{\sum_{j \in \mathcal{J}_{mt}} s_{ijmt} (v(q_{jt}) - \mathcal{E}[v(q)|\psi(q_{jt}), \psi])}_{\text{ex post correction}} \right), \quad (9)$$

where  $\delta_{ijmt}$  depends implicitly on the realization of quality and scoring policy. As the regulator evaluates (9) taking expectations over realized qualities, maximizing consumers' surplus consists of maximizing their ex ante utility from enrollment net of the correlation between choices and expectational errors.

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.

#### 5.4. Discussion

The model makes three 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.<sup>27</sup> Supplemental Appendix 2.3 examines this heterogeneity in the data, showing little evidence of meaningful differences across observable groups. Additionally, Supplemental Appendix 2.4 shows that the modeled heterogeneity in WTP is sufficient to generate over- and underprovision of quality and thus capture fundamental market frictions. In Supplemental Appendix 3.5, I discuss two robustness exercises that speak to the role of preference heterogeneity in scoring design.

The second 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 3 and 4 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. 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.

Finally, the analysis assumes enrollees' well-documented inertial enrollment behavior (Nosal (2011), Aizawa and Kim (2018)) is the product of heterogeneous cohort preferences rather than inattention or switching costs. Explicitly modeling inertia as inattention and allowing the regulator to force active choices would allow scores to coordinate more consumers. This would enhance their regulatory value and lead to larger welfare gains from optimal redesign than those documented here. Separately identifying inertia from systematic unobserved preferences, however, is a known challenge (Pakes, Porter, Shepard, and Calder-Wang (2022)).

<sup>27</sup>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.

## 6. IDENTIFICATION AND ESTIMATION

## 6.1. 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 (5) into their mean ( $\alpha$ ,  $\beta$ ) 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,  $\delta_{jmt}$ . This transformation has three unknown components: preference heterogeneity ( $\tilde{\alpha}_i$ ,  $\tilde{\beta}_i$ ), demographic preferences ( $\lambda^l$ ), and plan-market-year fixed effects ( $\delta$ ). Collecting these in a vector  $\boldsymbol{\vartheta}$ , the first stage solves

$$\max_{\boldsymbol{\vartheta}} \underbrace{\sum_t \sum_i w_{it} \sum_{j \in \mathcal{J}_m(i)t} y_{ijmt} \ln(s_{ijmt}(\boldsymbol{\vartheta}))}_{\text{weighted log-likelihood}} \quad \text{s.t.} \quad \underbrace{s_{jmt}^* = \sum_i w_{it} s_{ijmt}(\boldsymbol{\vartheta})}_{\text{share matching}} \quad \forall j, m, t, \quad (10)$$

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, Levinsohn, and Pakes (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 ( $\xi_{jmt}$ ) into systematic taste for MA in each market ( $\mathbb{d}_{mt}$ ), systematic preferences for the contract ( $\tilde{\eta}_{c(j)}$ ), and all residual unobserved preference ( $\tilde{\xi}_{jmt}$ ):

$$\begin{aligned} \hat{\delta}_{jmt} = & \underbrace{\alpha p_{jmt}^{\text{total}}}_{\text{premium}} + \underbrace{\beta b_{jmt}}_{\text{benefits}} + \underbrace{\lambda^{a'} a_{jmt}}_{\text{plan attributes}} \\ & + \underbrace{\mathcal{E}[v(q)|r_{jt}, \psi_t]}_{\text{quality}} + \underbrace{\tilde{\eta}_{c(j)}}_{\text{contract FE}} + \underbrace{\mathbb{d}_{mt}}_{\text{market-year FE}} + \tilde{\xi}_{jmt}. \end{aligned} \quad (11)$$

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.<sup>28</sup>

<sup>28</sup>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.

Consumers' unobserved preferences also influence firms' investments and, thus, scores. We can view consumers' preferences for quality and systematic preferences for contracts as a single endogenous contract-year fixed effect  $\eta_{c(j)t} = \mathcal{E}[v(q)|r_{jt}, \psi_t] + \bar{\eta}_{c(j)}$ . 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  $\omega_{kt}$  is the contribution of category  $k$  in year  $t$ . The first ratio,  $\omega_{kt}/\omega_{k't}$ , 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, 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.<sup>29</sup> 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.

Supplemental Appendix 2.5 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.

#### 6.1.1. *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—Informed Choice:** *Consumers know  $\psi_t(\cdot)$  and use scores and Bayes's rule to update a continuous prior density  $f: \mathcal{Q} \rightarrow \mathbb{R}_+$ , with compact and connected support.*

This assumption is common 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 Sfeekas (2008), Barahona, Otero, and Otero (2023)).<sup>30</sup> Crucially, in both cases, the econometrician knows how consumers interpret scores and can rely on their variation. In addition, consumers' knowledge of the scoring

<sup>29</sup>Even if all combinations were included, it would not fully predict a contract's rating as cutoffs vary.

<sup>30</sup>Some allow for parametric bias based on additional data, such as external surveys.



rule allows the regulator to shape their beliefs, which gives additional power to the scoring policy.<sup>31</sup>

Informed choice gains power once combined with variation in scores rich enough to match consumers' preference structure.<sup>32</sup> In Supplemental Appendix 2.6, 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. 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**—Preferences and Design Variation: *Quality preferences are linear, that is,  $v(q) = \gamma'q$ , and  $\psi_t$  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 consumers' preferences over quality are linear and that scores will continue to vary within a set, including, but not limited to, the type of designs observed in the data. It does not require the number of partitions to grow with the sample or entail complex aggregation rules (i.e., boundaries). The key identification result, proven in the Appendix, follows.

**PROPOSITION 1**—Quality Beliefs and Preference Identification: *Let Assumptions 1 and 2 hold; then  $(\gamma, f(\cdot))$  are identified.*

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.<sup>33</sup> 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  $\gamma$  within  $(8/9, 8/7)$ . Scoring variation produces new intervals for  $\gamma$ , intersecting and shrinking the identified set down to a point. This process also bounds posterior beliefs and, thus, priors.

I estimate quality preferences (now a vector  $\gamma$ ) and prior beliefs ( $f(\cdot)$ ) using a non-parametric minimum distance estimator. To remove any systematic contract preferences, I only leverage time-series variation in preferences for contract years,  $\eta_{c(j)t}$ . The estimator is

$$\min_{\gamma, \xi} \sum_{c(j)} \sum_t \sum_{\tau > t} (\Delta_t^\tau(\eta_{c(j)t} - \gamma' \mathcal{E}[q|r_{c(j)t}, \psi_t; \xi]))^2, \quad (12)$$

<sup>31</sup>Alternatively, rational expectations would imply consumers know the scoring rule, firms' costs, and investment risk well enough to predict quality changes. It would allow the regulator to control quality without informational losses, rendering informational policies stronger than those considered here.

<sup>32</sup>Scores imply lotteries over quality at different prices. If we observed consumers' preferences over all such lotteries, their subjective preferences and beliefs about quality would be identified without further meaningful structure (Anscombe and Aumann (1963)).

<sup>33</sup>The result does not depend on the logit structure. The full support assumption on scoring design is valuable for identifying the corners of prior beliefs, yet the argument is not at the limit: Variation in design provides meaningful identifying restrictions even if all partitions have positive measures.

TABLE II  
DEMAND ESTIMATES.

Panel A:	Premium ( $\alpha_i$ )		Benefits ( $\beta_i$ )	
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 ∈ [70, 75)	0.038	(0.060)	0.137	(0.053)
Age group ∈ [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 ( $\lambda^a$ )			Panel C: Quality Preferences ( $\gamma$ )		
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	36,447	Log likelihood	−5.403	Mean premium elasticity ( $p^C > 0$ )	−0.968
---	--------	----------------	--------	---------------------------------------	--------

Note: Panel A presents consumers’ estimated preferences for premiums and benefits, measured in thousands of dollars per year. The normalized group (i.e., mean preferences) 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 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 estimated quality preferences. Dividing the estimates by the absolute value of premium preferences results in consumers’ WTP for maximum quality in each dimension for a year, in thousands of dollars. Log likelihood is adjusted by MCBS sampling weights. Standard errors are homoscedastic and corrected for multi-stage estimation using the delta method. Demographic preferences ( $\lambda^l$ ) are shown in Supplemental Appendix Table 4. Supplemental Appendix Table 5 shows the effects of the instruments on these estimates.

where  $\Delta_t^\tau x_t \equiv x_t - x_{t-\tau}$  is the time difference operator and  $\zeta$  corresponds to the coefficients of a Fourier series expansion of the common prior  $f(\cdot)$ .<sup>34</sup>

6.1.2. Estimates

Panel A of Table II 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.<sup>35</sup> Higher-income and

<sup>34</sup>This step does not affect other estimates and can be safely disregarded when considering the assumption of policy-ignorance in Supplemental Appendix 4.

<sup>35</sup>The discrepancy with the findings of Abaluck and Gruber (2011) are, in part, due to \$1 in benefits translating to less than a \$1 reduction in expected spending. I discuss this further in Supplemental Appendix 2.8.

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.<sup>36</sup> 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 on how the regulation distorts firms' perceived elasticity in Supplemental Appendix 2.1.<sup>37</sup> As I will show later, this elasticity implies reasonable markups for firms in this market.

Panel B of Table II 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. Supplemental 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 II presents consumers' estimated 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 substantially less value on insurers facilitating preventive care or monitoring their health. These estimates imply that a low-income, low-predicted spending, low-risk score male aged 60 to 75 would be willing to pay \$4036 a year to access the highest possible quality of Medical Outcome but only \$1614 for the highest Intermediate Outcome quality. Mapping the estimates to the data, consumers are willing to pay \$12,004 for the median quality plan, roughly 24% more than this plan's revenue per member. 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.

### 6.1.3. *Informational Losses*

Consumers value quality, which they cannot observe. The estimates indicate that the average consumer loses \$199.3 in surplus due to incomplete information, equivalent to a third of a year's premiums. The losses stem from two 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.

<sup>36</sup>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.

<sup>37</sup>This is the elasticity relevant for firms' pricing decisions. 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  $-2.6$ , which would imply excessive price elasticities and negligible firm markups under this model.

Second, *across scores*, misalignment between consumers' preferences for quality categories and their relative 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 losses into these factors reveals that 94.5% stems from across-score frictions.<sup>38</sup> Within-score frictions are limited by firms' incentives to target the lower boundaries of scores.

#### 6.1.4. Selection

There is extensive literature studying selection into MA (Newhouse and McGuire (2014), Brown, Duggan, Kuziemko, and Woolston (2014), McGuire and Newhouse (2018)). This article's demand model captures an additional margin of selection based on attracting profitable consumers with quality (Glazer and McGuire (2006), Glazer, McGuire, Cao, and Zaslavsky (2008)). Supplemental Appendix Figure 1(a) shows consumers' WTP for quality decreases in their risk scores and increases in their predicted spending.<sup>39</sup> As higher-scoring enrollees contribute more to profits due to risk adjustment, firms are incentivized to attract high-risk scores with lower quality. Therefore, risk adjustment contributes to quality underprovision. Supplemental Appendix Figure 1(b), 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)), Supplemental Appendix 2.10 provides evidence that manipulation and selection are unlikely to play an important role in the scoring design problem.

### 6.2. Supply

#### 6.2.1. Insurance Marginal Costs

Insurers' pricing first-order optimality condition equates marginal revenue with marginal costs.<sup>40</sup> Since revenue depends only on observed demands, prices, and estimated elasticities, this condition can be used to recover the marginal cost parameters ( $\theta^c$ ). Assuming marginal costs are linear, the resulting condition is

$$\underbrace{p_f + R(p_f, z_f)}_{\text{revenue per consumer}} + \underbrace{(\nabla \tilde{D}_f)^{-1} (I + \nabla R_f(p_f, z_f)) \tilde{D}_f}_{-\text{profit margin}} = \underbrace{\theta_q^c q_f + \theta_a^c a_f + c_f}_{\text{marginal cost} = C(q_f, a_f, \theta^c)}, \quad (13)$$

where gradients are all with respect to the vector of prices  $p_f$ , and  $\tilde{D}_f$  is the risk-adjusted demand vector. On the right-hand side, I have decomposed the firm's marginal cost into its quality components ( $q_f$ ), systematic observable components ( $a_f$ ), and residual ( $c_f$ ).

Variations in demand and regulation identify marginal costs. I assume that the residual cost variation in  $q_f$  conditional on contract identity and year is unsystematic and unknown

<sup>38</sup>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.

<sup>39</sup>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. Consumers' WTP for coverage, as determined by the benefit level, is positively correlated with spending as expected.

<sup>40</sup>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.

TABLE III  
QUALITY'S INSURANCE COSTS, INVESTMENT COSTS, AND MARGINAL WELFARE.

Term	Access	Intermediate	Outcome	Patient	Process
<b>Panel A: Insurance cost</b>					
Linear ( $\theta_q^c$ )	30.077 (16.937)	104.038 (12.883)	15.903 (3.886)	-215.450 (57.988)	-176.811 (28.043)
<b>Panel B: Investment cost</b>					
Common linear ( $\mu_k$ )	1.514 (0.177)	-0.004 (0.182)	0.545 (0.208)	-2.595 (0.480)	3.010 (0.374)
PCP rate ( $\bar{\mu}_k$ )	-0.006 (0.001)	-0.005 (0.001)	-0.013 (0.001)	0.002 (0.003)	-0.007 (0.003)
Quadratic ( $\mu'_k$ )	-1.366 (0.197)	1.809 (0.312)	3.071 (0.472)	9.109 (0.0.614)	-0.475 (0.471)
<b>Panel C: Marginal welfare</b>					
Per contract-year	62.376 [3.9, 73.3]	17.647 [1.6, 25.1]	84.900 [8.4, 107.6]	65.287 [0.7, 102.3]	65.225 [2.5, 78.4]

*Note:* Panel A presents the marginal insurance cost coefficients associated with plan quality ( $\theta_q^c$ ) in dollars per unit-risk member month. The estimates for additional contract benefits ( $\theta_a^c$ ) are shown in Supplemental Appendix Table 6. The regression includes fixed effect for plan types (HMO, PPO, Regional plans, and PFFS), county, year, and contract identifiers. Standard errors in parentheses are heteroscedasticity robust and corrected for two-step estimation following [Murphy and Topel \(1985\)](#).  $N = 28,966$ ,  $R^2 = 0.529$ . Panel B presents the estimated annual investment cost parameters in millions of dollars per hundred thousand Medicare beneficiaries. The regression includes spillover components ( $\mu''_{k,k'}$ ), shown in the [Appendix Table 7](#), and linear terms for the interaction between the top six firms' identities and quality dimensions. The mean PCP rate is 83.1 per hundred thousand individuals.  $N = 7684$ . Standard errors are homoscedastic and unadjusted for multi-step estimation; see Supplemental Appendix II.N. Panel C shows the average derivative of total welfare ( $TW(\psi, 1, 1)$ ) with respect to each contract's quality in each dimension, in millions of dollars per year. The interquartile range is provided in brackets.

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 III](#) presents estimates of  $\theta_q^c$  when  $\mathbf{a}_f$  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.

### 6.2.2. 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 mapping  $q_{ck} = \Phi_k(x_{ck} + \epsilon_{ck})$ , where  $\Phi_k(\cdot)$  is an increasing function mapping the real line to the domain of quality dimension  $k$  and  $\epsilon_{ck}$  is the investment shock.<sup>41</sup> Under certain regularity conditions, the distribution of  $\epsilon_{ck}$  can be estimated nonparametrically using standard deconvolution results given observed quality realizations (Schennach (2016)). Supplemental Appendix 2.11 details the formal identification and nonparametric estimation procedure for risk.

Given the identified distributions of investment risk and quality, we can evaluate firms' investment optimality conditions in expectations. Formally, the first-order condition of investment for firm  $f$  in category  $k$  in year  $t$  equates marginal revenue ( $\frac{\partial}{\partial x_{ckt}} \mathbb{E}[V_f(\mathbf{q}_f, \mathbf{q}_{-f}, \psi_t) | \mathbf{x}_{ft}]$ ) with marginal investment cost ( $\frac{\partial I_f(\mathbf{x}_{ft})}{\partial x_{ckt}}$ ). Given observed quality  $\mathbf{q}_{ft}$ , we can decompose the marginal revenue into its conditional mean and variance, resulting in the condition<sup>42</sup>

$$\mathbb{E}\left[\frac{\partial}{\partial x_{ckt}} \mathbb{E}[V_f(\mathbf{q}_f, \mathbf{q}_{-f}, \psi_t) | \mathbf{x}_{ft}] | \mathbf{q}_{ft}\right] = \frac{\partial I_f(\mathbf{x}_{ft})}{\partial x_{ckt}} + \nu_{ckt}, \quad \mathbb{E}[\nu_{ckt} | \mathbf{q}_{ft}] = 0. \quad (14)$$

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 (14) is a regression equation. To operationalize it, I model firms' investment costs as

$$I_f(\mathbf{x}_{ft}) = \sum_{c \in \mathcal{C}_{ft}, k \in \mathcal{K}} \underbrace{M_{ct}}_{\text{population}} \left( \underbrace{\mu_{ckt} \tilde{x}_{ckt} + \frac{\mu'_k}{2} \tilde{x}_{ckt}^2}_{\text{category-specific cost}} + \underbrace{\sum_{k' \neq k} \frac{\mu''_{k,k'}}{2} \tilde{x}_{ckt} \tilde{x}_{ck't}}_{\text{cross-category spillovers}} \right), \quad (15)$$

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 parentheses are the category-specific quadratic costs, where I parameterize the first as  $\mu_{ckt} = \mu_k + \tilde{\mu}_{k,f(c)} + \bar{\mu}_k \text{PCP}_{ct} + \epsilon_{ckt}^\mu$ . In this expression, the first coefficient captures common investment costs, while the second captures firms' cost advantages. The

<sup>41</sup>The definition of  $\Phi_k$  is arbitrary, as  $x_{ck}$  is a modeling device. I take  $\Phi_k(x) = \Phi(x)(1 - q_k)$ , where  $\Phi(\cdot)$  is the standard normal CDF, and  $q_k$  is the minimum of  $q_k$  in the data, which is constrained above zero by minimum quality regulation.

<sup>42</sup>Supplemental Appendix 2.12 describes how I estimate firms' rational expectations about rivals' actions, which are necessary to evaluate this expectation. Supplemental Appendix 2.13 shows that the left-hand side of this expression is a function of only identified distributions and has an analytical expression.



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,  $\epsilon_{\text{ckt}}^\mu$  is an i.i.d. mean-zero unobserved shock to investment costs. The final term in Equation (15) 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 (14) and the investment cost function of Equation (15) to evaluate the implied moment condition ( $\mathbb{E}[\nu_{\text{ckt}}|q_{ft}] = 0$ ) using observed quality. This involves replacing  $\tilde{x}_{\text{ckt}}$  in (15) with its closest observable analog,  $\Phi_k^{-1}(q_{\text{ckt}}) - \Phi_k^{-1}(q_{kt})$ , where  $q_{kt}$  is the minimum quality for dimension  $k$  in year  $t$ . This replacement and the unobserved cost shock  $\epsilon_{\text{ckt}}^\mu$  create an endogeneity problem: The realized quality will tend to be 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  $c$ .<sup>43</sup> 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.<sup>44</sup>

Panel B of Table III 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

<sup>43</sup>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  $\tilde{\xi}_{jmt}$ .

<sup>44</sup>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.

across providers of screening and labs. It is also reasonable that it does not affect patient satisfaction, as the additional PCP might not be better than those found in tighter markets.

Supplemental 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 (15) is normalized such that firms investing at the minimum level ( $x_{kt}$ ) pay an investment cost of zero.

### 6.2.3. 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 III 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. Supplemental Appendix Table 8 shows the results of regressing the marginal welfare value of quality on HHI, the category's contribution in the scoring 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.

## 7. 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 the Supplemental Material (Vatter (2025)).

### 7.1. Approach

The designer seeks to maximize total welfare,  $TW(\psi, \rho^F, \rho^G)$  in Equation (8), by choosing a scoring rule  $\psi \in \Psi$ . In choosing, she recognizes the scores' effect on equilibrium investments, prices, beliefs, and enrollment. This presents her with a trade-off between information and efficiency: 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.

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.

First, I show that monotone partitional scores are a composition of a polynomial *aggregator*, aggregating multidimensional quality into an index, and a *cutoff* function, which partitions the index into scores (see Supplemental Appendix 3.1). Therefore, the designer can solve the problem by finding the best score 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, Maschler, and Stearns \(1995\)](#), who noted 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 ( $\mathcal{E}[\gamma'q|r, \psi]$ ), and marginal quality costs ( $\theta'q$ ). 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 scores are discontinuous and firms compete over multidimensional quality, 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 ensures convergence would fail if such equilibrium is not feasible and unique. In addition, to avoid local optima in the regulator's objective and find the global maximum, I use the algorithm of [Malherbe and Vayatis \(2017\)](#), which provides convergence guarantees for Lipschitz continuous functions of multiple bounded arguments. Supplemental Appendix 3.2 offers further details on implementation.

The following results focus on markets included in the MCBS 2015 data, covering nearly 22 million beneficiaries. Given mixed evidence on the extent of selection into MA ([Newhouse, Price, McWilliams, Hsu, and McGuire \(2015\)](#)), I omit subsidy spending from the main analysis, setting  $\rho^F = 1$  and  $\rho^G = 0$  in Equation (8). I discuss different objectives in Section 7.5. The results are under the assumption of informed choice (Assumption 1); Supplemental Appendix 4 presents results for policy-ignorant consumers, showing that the main lessons from the analysis hold.

The first row of Table IV 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, full-information welfare numbers should be considered cautiously as scores help consumers compare across options, reducing enrollment complexity. As the data contain no meaningful variation in choice or scoring complexity, the analysis addresses this gap by restricting counterfactual scoring policies to ones at most as complex as the status quo, keeping the net change in complexity 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.

TABLE IV  
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	$\Delta$ Consumer Surplus	$\Delta$ Firm Profits	$\Delta$ Total Welfare	$\Delta$ Info.	$\Delta$ 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	

Note: 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. 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 Supplemental Appendix 3 for adjusted numbers.

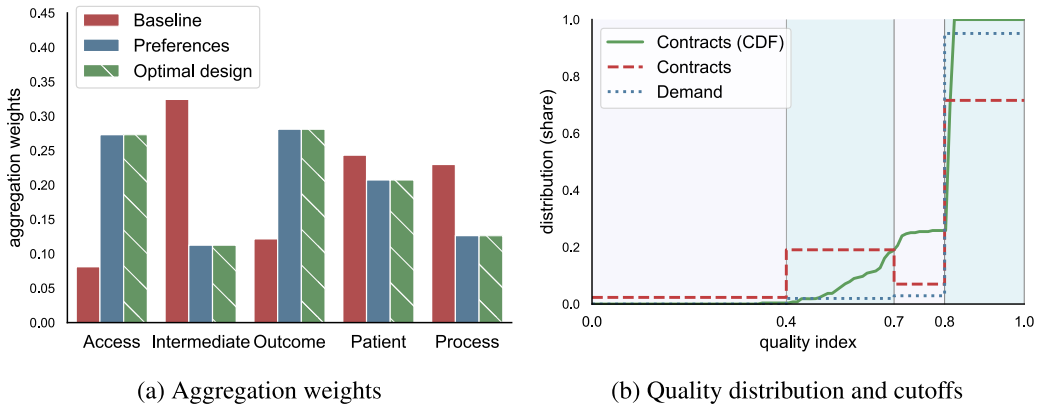


FIGURE 4.—Optimal design. *Note:* The optimal design comprises aggregation weights that reduce multi-dimensional quality into a single index and cutoffs that partition the index into scoring regions. Figure (a) compares the optimal aggregation weights (hatched) with CMS’s average aggregator for 2015 (leftmost bar) and consumers’ preferences normalized to a unit sum (middle bars). Figure (b) shows the scoring cutoff along the quality index, with segments indicating different scores. The solid line shows the equilibrium cumulative distribution of contracts across scores, the dashed line shows the share of contracts per score, and the dotted line shows 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.

## 7.2. Optimal Design

Figure 4 shows the optimal monotone partitional design, and the first row of Panel A in Table IV 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.

### 7.2.1. 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 4(b) 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

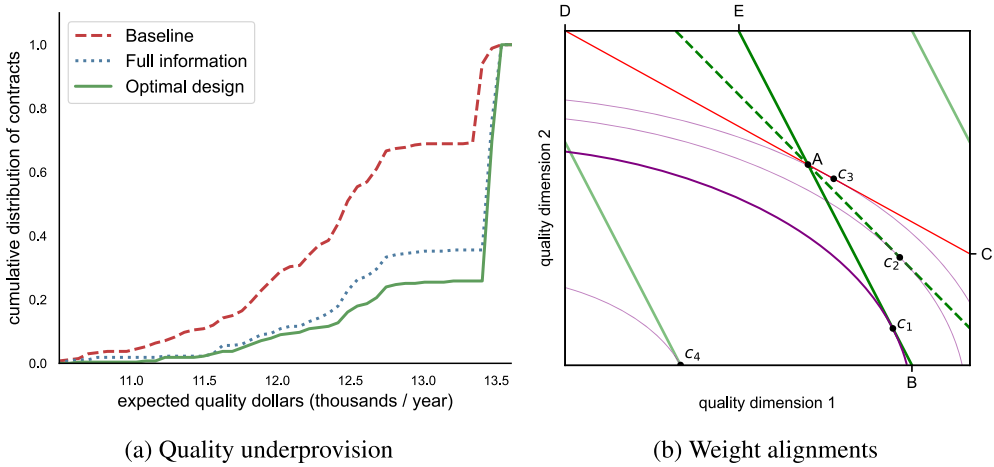


FIGURE 5.—Scoring design mechanisms. *Note:* Figure (a) plots the distribution of quality in the baseline, full-information benchmark, and optimal design. Quality is measured according to consumers’ WTP, in thousands of dollars per year. The distribution does not match that of Figure 4 due to differences 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 in  $ABC$ . The dashed line represents a potential redesign. Concave lines are a firm’s isocost curve under different total investment levels.

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 2. Figure 5(a) 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.

### 7.2.2. 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 Zapechelnuk (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 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.



### 7.2.3. 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 5(b) 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 5.2. 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 5(b), point  $c_1$  marks the tangency of a firm's concave isocost curve with the linear scoring threshold, 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 5(b) since products in the triangle  $\triangle DEA$  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 IV, 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 5(b), 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.

### 7.3. Welfare

The third row of Table IV 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 Supplemental 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 by as much as 16.6%, while the overall average

change is 4% as shown in Table IV. 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 Supplemental Appendix 3.4.<sup>45</sup>

### 7.3.1. *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  $(x^*, p^*, \Psi^*)$  denote the optimal investment, prices, and scoring policy, and  $(x^0, p^0, \Psi^0)$  the baseline, the compensating variation value of quality is  $CV_q = TW(x^*, p^*, \Psi^*) - TW(x^0, p^*, \Psi^*)$ . 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 IV 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. 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 IV 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 2, consumers in a scored market can behave as if fully informed, even if they cannot

<sup>45</sup>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.

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.

### 7.3.2. *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 (see Supplemental Appendix Figure 3). First, I find the optimal certification scheme that preserves CMS's average aggregation weight for 2015. The new design—whose equilibrium impact is shown in the third row of Panel A in Table IV—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 IV. 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 IV. 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. 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.

### 7.4. *Private Information and Top-Revealing Designs*

The data required to identify insurers' costs are available only after the market is realized and rely, in part, on the regulator's experimentation. Therefore, in principle, the regulator might have uncertainty about insurers' costs when designing the policy. To capture this scenario, I model the uncertainty as five independent mean-zero normal distributions over the heterogeneous linear investment cost terms ( $\mu_{\text{ckt}}$ ), one for each quality category. I set 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 on certification designs and take ten draws from the cost distributions. The first row of Panel B in Table IV shows the outcomes, while Supplemental Appendix Figure 4 shows the 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 IV 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) proved 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.<sup>46</sup> This policy is not a coarse monotone partitional design, as covered by the main solver. However, modifying the exploration strategy to include it is simple. Supplemental Appendix Figure 4 shows the optimal top-revealing certification, and the second row of Panel B in Table IV shows its welfare impact. Top-revelation increases welfare beyond 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.

Top-revealing designs can also offer improvements when the regulator is fully informed of insurers' costs, as in the main analysis. Supplemental Appendix Figure 5 shows the optimal top-revealing design, and the last row of Panel A in Table IV shows the resulting outcomes.<sup>47</sup> 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. Relative to the baseline design, the aggregator is better aligned with consumers' preferences to reduce 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 adopting simpler approximations to the theoretical optimum.

### 7.5. *Alternative Regulatory Objectives*

Panel C of Table IV and Supplemental 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 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

<sup>46</sup>Zapechelnyuk (2020) considered 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.

<sup>47</sup>This design was optimized subject to the constraint of at most nine scores and a linear aggregator.

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.

### 7.6. *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 toward 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 Supplemental 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.<sup>48</sup> Scores are a poor nudging mechanism as it is inherently costly to steer consumers with information they do not value.

An alternative explanation for CMS's design is given in Supplemental Appendix IV, which considers a regulator that treats consumers as naive, with beliefs about quality independent and invariant to the scoring design. As shown in Section 4, this naivety is rejected by the data. However, if the regulator believes it and is extremely averse to misrepresenting consumers' preferences, the status quo system outperforms the best simple scoring design. Given Medicare's delicate political and social role, these findings are, perhaps, reasonable.

### 7.7. *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

<sup>48</sup>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.

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 high-value outside options. This finding also implies a contradiction between efforts to regulate and disclose quality, as neither quality nor consumers' information is monotonic in the *ex ante* (i.e., without equilibrium effects) informativeness of scores. More granular systems can worsen 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 for 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 plaguing various disclosure policies (Feng Lu (2012), Reynaert and Sallee (2021)). The results show scores can align firms' incentives with regulatory objectives and allow heterogeneous firms to reach high-quality production through various paths, stimulating a variety of products not permitted by stricter minimum quality standard policies. 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)).

## 8. 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.



## APPENDIX A: PROOF OF PROPOSITION 1

The following proof is for scalar quality ( $|\mathcal{Q}| = 1$ ). The extension to multiple dimensions is relegated to Section 2 of the Supplemental Material (Vatter (2025)). Throughout, I assume score-year fixed effect  $\eta_{rt} = \gamma' \mathcal{E}[q|\psi(q) = r] + \bar{\eta}$  are identified up to a constant  $\bar{\eta}$ . In MA, this corresponds to the mean valuation for MA relative to TM. The proof depends on the following preliminary lemma.

LEMMA 2: *Let  $f, g$  be two distinct, continuous, strictly positive densities supported on  $[0, 1]$ . Then, there exists  $\underline{x} < \tilde{x} < \bar{x} \in [0, 1]$  such that either  $\mathbb{E}_f[x|x \in (\underline{x}, \tilde{x})] \geq \mathbb{E}_g[x|x \in (\underline{x}, \tilde{x})]$  and  $\mathbb{E}_f[x|x \in (\tilde{x}, \bar{x})] \leq \mathbb{E}_g[x|x \in (\tilde{x}, \bar{x})]$  with one of the inequalities strict, or the analogous statement holds with the roles of  $f, g$  reversed. Also, there exists another  $\underline{x} < \bar{x} \in [0, 1]$  such that  $\mathbb{E}_f[x|x \in (\underline{x}, \bar{x})] = \mathbb{E}_g[x|x \in (\underline{x}, \bar{x})]$ .*

PROOF: By continuity and common support,  $f$  and  $g$  cross at  $\tilde{x} \in (0, 1)$ . By continuity,  $\exists \epsilon > 0$  such that, wlog,  $f(x) > g(x) \forall x \in (\tilde{x}, \tilde{x} + \epsilon)$  and  $f(x) \leq g(x)$  in  $(\tilde{x} - \epsilon, \tilde{x})$ . Define  $h_f(x, \epsilon) = f(x)/(F(\tilde{x} + \epsilon) - F(\tilde{x}))$  and analogously for  $g$ , where  $F$  is the cumulative of  $f$  and  $G$  that of  $g$ . Note that  $\forall \bar{\epsilon} \in (0, \epsilon)$  we have that  $h_f(\tilde{x}, \bar{\epsilon}) < h_g(\tilde{x}, \bar{\epsilon})$ , and that both  $h_f(\cdot, \bar{\epsilon})$  and  $h_g(\cdot, \bar{\epsilon})$  are continuous densities integrating to one within  $(0, \bar{\epsilon})$  and therefore intersect at an interior point. Pick  $\bar{\epsilon} \in (0, \epsilon)$  such that  $h_f(\cdot, \bar{\epsilon})$  and  $h_g(\cdot, \bar{\epsilon})$  intersect only once at a point  $\hat{x}$ . Denote  $\bar{x} = \tilde{x} + \bar{\epsilon}$ . Then we have that

$$\begin{aligned} & \mathbb{E}_f[x|x \in (\tilde{x}, \bar{x})] - \mathbb{E}_g[x|x \in (\tilde{x}, \bar{x})] \\ &= \int_{\tilde{x}}^{\bar{x}} (h_f(v, \bar{\epsilon}) - h_g(v, \bar{\epsilon})) v dv \\ &= \int_{\tilde{x}}^{\hat{x}} \underbrace{(h_f(v, \bar{\epsilon}) - h_g(v, \bar{\epsilon}))}_{<0} v dv + \int_{\hat{x}}^{\bar{x}} \underbrace{(h_f(v, \bar{\epsilon}) - h_g(v, \bar{\epsilon}))}_{>0} v dv \\ &> \hat{x} \int_{\tilde{x}}^{\hat{x}} (h_f(v, \bar{\epsilon}) - h_g(v, \bar{\epsilon})) dv + \hat{x} \int_{\hat{x}}^{\bar{x}} (h_f(v, \bar{\epsilon}) - h_g(v, \bar{\epsilon})) dv = 0. \end{aligned}$$

This proves the first inequality. The proof for the second is analogous applied to  $(\tilde{x} - \epsilon, \tilde{x})$ . The third statement follows from the intermediate value theorem applied to  $w(\lambda) = \mathbb{E}_f[x|x \in (\underline{x}, \tilde{x} + \lambda\bar{\epsilon})] - \mathbb{E}_g[x|x \in (\underline{x}, \tilde{x} + \lambda\bar{\epsilon})]$ , noting that  $w(1) > 0$ ,  $w(0) \leq 0$  and  $w(\cdot)$  is continuous. Q.E.D.

We can now state the proof of Proposition 1.

PROOF: By contradiction, suppose there exist two distinct  $(\gamma_0, f_0, \bar{\eta}_0)$ ,  $(\gamma_1, f_1, \bar{\eta}_1)$  in the identified set  $\mathcal{I}$ . By Lemma 2 and Assumption 2, there exist two monotone partitional designs  $\tilde{\psi}$  and  $\psi$  drawn with positive probability, such that: (1) for  $\tilde{\psi}$  there is a partition  $\tilde{r}$  where  $\mathbb{E}_{f_0}[q|\tilde{r}, \tilde{\psi}] = \mathbb{E}_{f_1}[q|\tilde{r}, \tilde{\psi}]$ ; (2) for  $\psi$  there are two partitions  $r, r'$  such that  $\mathbb{E}_{f_0}[q|r, \psi] < \mathbb{E}_{f_1}[q|r, \psi]$  and  $\mathbb{E}_{f_0}[q|r', \psi] \geq \mathbb{E}_{f_1}[q|r', \psi]$ , where the directions of the inequality are assumed without loss. Using this, we have that

$$\begin{aligned} & \gamma_0(\mathbb{E}_{f_0}[q|r, \psi] - \mathbb{E}_{f_0}[q|\tilde{r}, \tilde{\psi}]) \\ &= \eta_{rt} - \tilde{\eta}_{\tilde{r}} = \gamma_1(\mathbb{E}_{f_1}[q|r, \psi] - \mathbb{E}_{f_1}[q|\tilde{r}, \tilde{\psi}]) \implies \gamma_0 > \gamma_1, \end{aligned}$$

$$\begin{aligned} & \gamma_0(\mathbb{E}_{f_0}[q|r', \psi] - \mathbb{E}_{f_0}[q|\tilde{r}, \tilde{\psi}]) \\ &= \eta_{r't} - \tilde{\eta}_{\tilde{r}} = \gamma_1(\mathbb{E}_{f_1}[q|r', \psi] - \mathbb{E}_{f_1}[q|\tilde{r}, \tilde{\psi}]) \implies \gamma_0 \leq \gamma_1, \end{aligned}$$

contradicting that  $(\gamma_0, f_0, \bar{\eta}_0)$  and  $(\gamma_1, f_1, \bar{\eta}_1)$  are in the identified set.  $\mathcal{I}$  is a singleton.  
*Q.E.D.*

## REFERENCES

- ABALUCK, JASON, AND JONATHAN GRUBER (2011): “Choice Inconsistencies Among the Elderly: Evidence From Plan Choice in the Medicare Part D Program,” *American Economic Review*, 101 (4), 1180–1210. [0962, 0973, 0980]
- ABALUCK, JASON, MAURICIO CACERES BRAVO, PETER HULL, AND AMANDA STARC (2021): “Mortality Effects and Choice Across Private Health Insurance Plans,” *The Quarterly Journal of Economics*, 136 (3), 1557–1610. [0960]
- AIZAWA, NAOKI, AND YOU SUK KIM (2018): “Advertising and Risk Selection in Health Insurance Markets,” *American Economic Review*, 108 (3), 828–867. [0962, 0967, 0973, 0976]
- ALBANO, GIAN LUIGI, AND ALESSANDRO LIZZERI (2001): “Strategic Certification and Provision of Quality,” *International Economic Review*, 42 (1), 267–283. [0961]
- ALÉ-CHILET, JORGE, AND SARAH MOSHARY (2022): “Beyond Consumer Switching: Supply Responses to Food Packaging and Advertising Regulations,” *Marketing Science*, 41 (2), 243–270. [0961]
- ALLCOTT, HUNT (2011): “Consumers’ Perceptions and Misperceptions of Energy Costs,” *American Economic Review*, 101, 98–104. [0975]
- ALLENDE, CLAUDIA, FRANCISCO GALLEGO, AND CHRISTOPHER NEILSON (2019): “Approximating the Equilibrium Effects of Informed School Choice,” Working Paper, Princeton University, Princeton NJ. [0962]
- ANGRIST, JOSHUA D., AND JONATHAN GURRYAN (2008): “Does Teacher Testing Raise Teacher Quality? Evidence From State Certification Requirements,” *Economics of Education Review*, 27 (5), 483–503. [0962]
- ANSCOMBE, FRANCIS JOHN, AND ROBERT JOHN AUMANN (1963): “A Definition of Subjective Probability,” *The Annals of Mathematical Statistics*, 34 (1), 199–205. [0979]
- ARAYA, SEBASTIAN, ANDRES ELBERG, CARLOS NOTON, AND DANIEL SCHWARTZ (2018): “Identifying Food Labeling Effects on Consumer Behavior,” *SSRN Electronic Journal*. [0961]
- ATAL, JUAN PABLO, JOSÉ IGNACIO CUESTA, AND MORTEN SÆTHRE (2022): “Quality Regulation and Competition: Evidence From Pharmaceutical Markets,” National Bureau of Economic Research Working Paper Series. [0962]
- AUMANN, ROBERT J., MICHAEL MASCHLER, AND RICHARD E. STEARNS (1995): *Repeated Games With Incomplete Information*. MIT press. [0987]
- BALL, IAN (2025): “Scoring Strategic Agents,” *American Economic Journal: Microeconomics*, 17 (1), 97–129. [0961]
- BARAHONA, NANO, CRISTÓBAL OTERO, AND SEBASTIÁN OTERO (2023): “Equilibrium Effects of Food Labeling Policies,” *Econometrica*, 91 (3), 839–868. [0961, 0962, 0973, 0978]
- BERRY, STEVEN, AND PHILIP HAILE (2020): “Nonparametric Identification of Differentiated Products Demand Using Micro Data,” National Bureau of Economic Research. [0978]
- BERRY, STEVEN, JAMES LEVINSOHN, AND ARIEL PAKES (1995): “Automobile Prices in Market Equilibrium,” *Econometrica*, 63 (4), 841. [0977]
- BERRY, STEVEN T. (1994): “Estimating Discrete-Choice Models of Product Differentiation,” *The RAND Journal of Economics*, 25 (2), 242. [0977]
- BLACKWELL, DAVID (1953): “Equivalent Comparisons of Experiments,” *The annals of mathematical statistics*, 265–272. [0959]
- BLATTNER, LAURA, JACOB HARTWIG, AND SCOTT NELSON (2022): “Information Design in Consumer Credit Markets,” Working paper. [0961]
- BOLESILAVSKY, RAPHAEL, AND KYUNGMIN KIM (2018): “Bayesian Persuasion and Moral Hazard,” *SSRN Electronic Journal*. [0962]
- BROWN, JASON, MARK DUGGAN, ILYANA KUZIEMKO, AND WILLIAM WOOLSTON (2014): “How Does Risk Selection Respond to Risk Adjustment? New Evidence From the Medicare Advantage Program,” *American Economic Review*, 104 (10), 3335–3364. [0982]
- BROWN, ZACH Y., AND JIHYE JEON (2024): “Endogenous Information and Simplifying Insurance Choice,” *Econometrica*, 92 (3), 881–911. [0962]
- CMS (2016): “Quality Strategy,” Technical report. [0960]

- COOPER, ZACK, STEPHEN GIBBONS, SIMON JONES, AND ALISTAIR MCGUIRE (2011): "Does Hospital Competition Save Lives? Evidence From the English NHS Patient Choice Reforms," *Economic Journal*, 121 (554), 228–260. [0962]
- CRAWFORD, GREGORY S., AND MATTHEW SHUM (2005): "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica*, 73 (4), 1137–1173. [0978]
- CRAWFORD, GREGORY S., OLEKSANDR SHCHERBAKOV, AND MATTHEW SHUM (2019): "Quality Overprovision in Cable Television Markets," *American Economic Review*, 109 (3), 956–995. [0960,0963]
- CURTO, VILSA, LIRAN EINAV, AMY FINKELSTEIN, JONATHAN LEVIN, AND JAY BHATTACHARYA (2019): "Health Care Spending and Utilization in Public and Private Medicare," *American Economic Journal: Applied Economics*, 11 (2), 302–332. [0968,0983]
- CURTO, VILSA, LIRAN EINAV, JONATHAN LEVIN, AND JAY BHATTACHARYA (2021): "Can Health Insurance Competition Work? Evidence From Medicare Advantage," *Journal of Political Economy*, 129 (2), 570–606. [0962,0965,0967,0973]
- CUTLER, DAVID M., ROBERT S. HUCKMAN, AND JONATHAN T. KOLSTAD (2010): "Input Constraints and the Efficiency of Entry: Lessons From Cardiac Surgery," *American Economic Journal: Economic Policy*, 2 (1), 51–76. [0962]
- DAFNY, LEEMORE, AND DAVID DRANOVE (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. [0966,0968]
- DAI, WEIJIA DAISY, GINGER JIN, JUNGMIN LEE, AND MICHAEL LUCA (2018): "Aggregation of Consumer Ratings: An Application to Yelp.com," *Quantitative Marketing and Economics*, 16 (3), 289–339. [0961]
- DARDEN, MICHAEL, AND IAN M. MCCARTHY (2015): "The Star Treatment: Estimating the Impact of Star Ratings on Medicare Advantage Enrollments," *Journal of Human Resources*, 50 (4), 980–1008. [0968]
- DECAROLIS, FRANCESCO, ANDREA GUGLIELMO, AND CLAVIN LUSCOMBE (2020): "Open Enrollment Periods and Plan Choices," *Health Economics*, 29 (7), 733–747. [0962]
- DECAROLIS, FRANCESCO, MARIA POLYAKOVA, AND STEPHEN P. RYAN (2020): "Subsidy Design in Privately Provided Social Insurance: Lessons From Medicare Part D," *Journal of Political Economy*, 128 (5), 1712–1752. [0962]
- DRANOVE, DAVID, AND GINGER ZHE JIN (2010): "Quality Disclosure and Certification: Theory and Practice," *Journal of Economic Literature*, 48 (4), 935–963. [0961]
- DRANOVE, DAVID, AND ANDREW SFEKAS (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. [0978]
- DWORCZAK, PIOTR, AND GIORGIO MARTINI (2019): "The Simple Economics of Optimal Persuasion," *Journal of Political Economy*, 127 (5), 1993–2048. [0979]
- ELFENBEIN, DANIEL W., RAYMOND FISMAN, AND BRIAN MCMANUS (2015): "Market Structure, Reputation, and the Value of Quality Certification," *American Economic Journal: Microeconomics*, 7 (4), 83–108. [0961]
- FENG LU, SUSAN (2012): "Multitasking, Information Disclosure, and Product Quality: Evidence From Nursing Homes," *Journal of Economics and Management Strategy*, 21 (3), 673–705. [0996]
- FIORETTI, MICHELE, AND HONGMING WANG (2021): "Performance Pay in Insurance Markets: Evidence From Medicare," *The Review of Economics and Statistics*, 1–45. [0982]
- FLEITAS, SEBASTIAN (2020): "Who Benefits When Inertia is Reduced? Competition, Quality and Returns to Skill in Health Care Markets," CEPR Discussion Paper. [0962]
- FRANK, RICHARD G., AND THOMAS G. MCGUIRE (2019): "Market Concentration and Potential Competition in Medicare Advantage," Issue brief (Commonwealth Fund), 2019 (February), 1–8. [0965]
- GAYNOR, MARTIN, RODRIGO MORENO-SERRA, AND CAROL PROPPER (2013): "Death by Market Power: Reform, Competition, and Patient Outcomes in the National Health Service," *American Economic Journal: Economic Policy*, 5 (4), 134–166. [0962]
- GERUSO, MICHAEL, AND TIMOTHY LAYTON (2020): "Upcoding: Evidence From Medicare on Squishy Risk Adjustment," *Journal of Political Economy*, 128, 984–1026. [0982]
- GLAZER, JACOB, AND THOMAS G. MCGUIRE (2006): "Optimal Quality Reporting in Markets for Health Plans," *Journal of Health Economics*, 25, 295–310. [0961,0982]
- GLAZER, JACOB, THOMAS G. MCGUIRE, ZHUN CAO, AND ALAN ZASLAVSKY (2008): "Using Global Ratings of Health Plans to Improve the Quality of Health Care," *Journal of Health Economics*, 27, 1182–1195. [0982]
- GOOLSBEE, AUSTAN, AND AMIL PETRIN (2004): "The Consumer Gains From Direct Broadcast Satellites and the Competition With Cable TV," *Econometrica*, 72 (2), 351–381. [0977]
- HANDEL, BEN, IGAL HENDEL, AND MICHAEL D. WHINSTON (2015): "Equilibria in Health Exchanges: Adverse Selection versus Reclassification Risk," *Econometrica*, 83 (4), 1261–1313. [0962]
- HANDEL, BENJAMIN R. (2013): "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, 103 (7), 2643–2682. [0962,0973]

- HANDEL, BENJAMIN R., AND JONATHAN T. KOLSTAD (2015): "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 105 (8), 2449–2500. [0962]
- HARBAUGH, RICK, AND ERIC RASMUSEN (2018): "Coarse Grades: Informing the Public by Withholding Information," *American Economic Journal: Microeconomics*, 10 (1), 210–235. [0961]
- HO, KATE, AND BENJAMIN HANDEL (2021): "The industrial organization of health care markets," in *Handbook of industrial organization*, Vol. 5, 521–614. (2021). [0962]
- HO, KATE, AND ROBIN S. LEE (2017): "Insurer Competition in Health Care Markets," *Econometrica*, 85 (2), 379–417. [0962]
- HO, KATE, JOSEPH HOGAN, AND FIONA SCOTT MORTON (2017): "The Impact of Consumer Inattention on Insurer Pricing in the Medicare Part D Program," *The RAND Journal of Economics*, 48 (4), 877–905. [0962]
- HOLMSTROM, BENGT, AND PAUL MILGROM (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *The Journal of Law, Economics, and Organization*, 7, 24–52. [0963]
- HOPENHAYN, HUGO, AND MARYAM SAEEDI (2019): "Optimal Ratings and Market Outcomes," NBER Working Paper Series, 1–39. [0961]
- HOUE, SÉBASTIEN (2018): "Bunching With the Stars: How Firms Respond to Environmental Certification," *SSRN Electronic Journal*. [0962]
- JIN, GINGER ZHE, AND PHILLIP LESLIE (2003): "The Effect of Information on Product Quality: Evidence From Restaurant Hygiene Grade Cards," *Quarterly Journal of Economics*, 118 (2), 409–451. [0962]
- JIN, GINGER ZHE, AND ALAN T. SORENSEN (2006): "Information and Consumer Choice: The Value of Publicized Health Plan Ratings," *Journal of Health Economics*, 25 (2), 248–275. [0961,0975]
- KAMENICA, EMIR (2019): "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 11 (1), 249–272. [0961]
- KAMENICA, EMIR, AND MATTHEW GENTZKOW (2011): "Bayesian Persuasion," *American Economic Review*, 101 (6), 2590–2615. [0978]
- KLEINER, MORRIS, AND EVAN SOLTAS (2019): "A Welfare Analysis of Occupational Licensing in U.S. States," National Bureau of Economic Research Working Paper Series. [0962]
- KOLOTILIN, ANTON, AND ANDRIY ZAPECHELNYUK (2019): "Persuasion Meets Delegation," arXiv preprint. arXiv:1902.02628. [0960,0990]
- KOLSTAD, JONATHAN T. (2013): "Information and Quality When Motivation is Intrinsic: Evidence From Surgeon Report Cards," *American Economic Review*, 103 (7), 2875–2910. [0962]
- LARSEN, BRADLEY, ZIAO JU, ADAM KAPOR, AND CHUAN YU (2020): "The Effect of Occupational Licensing Stringency on the Teacher Quality Distribution," National Bureau of Economic Research Working Paper Series. [0962]
- MALHERBE, CÉDRIC, AND NICOLAS VAYATIS (2017): "Global Optimization of Lipschitz Functions," in *34th International Conference on Machine Learning, ICML 2017*, Vol. 5, 3592–3601. (1972). [0987]
- MARONE, VICTORIA R., AND ADRIENNE SABETY (2022): "When Should There Be Vertical Choice in Health Insurance Markets?" *American Economic Review*, 112 (1), 304–342. [0962]
- MCGUIRE, THOMAS G., AND JOSEPH P. NEWHOUSE (2018): "Chapter 19—Medicare Advantage: Regulated Competition in the Shadow of a Public Option," Academic Press, 563–598. [0982]
- MCGUIRE, THOMAS G., JOSEPH P. NEWHOUSE, AND ANNA D. SINAICO (2011): "An Economic History of Medicare Part C," *Milbank Quarterly*, 89 (2), 289–332. [0965]
- MEDPAC (2018): "The Medicare Advantage Program: Status Report," Report to Congress: Medicare Payment Policy, 287–306. [0996]
- MEDPAC (2020): "The Medicare Advantage Program: Status Report," Report to Congress: Medicare Payment Policy, 287–306. [0996]
- MILLER, KEATON S., AMIL PETRIN, ROBERT TOWN, AND MICHAEL CHERNEW (2022): "Optimal Managed Competition Subsidies," NBER Working Paper Series. [0962,0967,0973,0981]
- MURPHY, KEVIN M., AND ROBERT H. TOPEL (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business & Economic Statistics*, 20 (1), 88–97. [0983]
- MUSSA, MICHAEL, AND SHERWIN ROSEN (1978): "Monopoly and Product Quality," *Journal of Economic Theory*, 18 (2), 301–317. [0976]
- NEWHOUSE, JOSEPH P., AND THOMAS G. MCGUIRE (2014): "How Successful is Medicare Advantage?" *Milbank Quarterly*, 92 (2), 351–394. [0982,0983]
- NEWHOUSE, JOSEPH P., MARY PRICE, J. MICHAEL MCWILLIAMS, JOHN HSU, AND THOMAS G. MCGUIRE (2015): "How Much Favorable Selection is Left in Medicare Advantage?" *American journal of health economics*, 1 (1), 1–26. [0987]
- NOSAL, KATHLEEN (2011): "Estimating Switching Costs for Medicare Advantage Plans," Working paper. [0976]

- PAKES, ARIEL, JACK PORTER, MARK SHEPARD, AND SOPHIE CALDER-WANG (2022): "Unobserved Heterogeneity, State Dependence, and Health Plan Choices," NBER Working paper. [0976]
- POLYAKOVA, MARIA (2016): "Regulation of Insurance With Adverse Selection and Switching Costs: Evidence From Medicare Part D," *American Economic Journal: Applied Economics*, 8 (3), 165–195. [0962]
- POPULATION HEALTH INSTITUTE, UNIVERSITY OF WISCONSIN (2024): "County Health Rankings & Roadmaps 2024," <https://www.countyhealthrankings.org/>. [0968]
- REID, RACHEL O., PARTHA DEB, BENJAMIN L. HOWELL, AND WILLIAM H. SHRANK (2013): "Plan Star Ratings and Enrollment," *Journal of the American Medical Association*, 309 (3), 267–274. [0968]
- REIMERS, IMKE, AND JOEL WALDFOGEL (2021): "Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings," *American Economic Review*, 111, 1944–1971. [0975]
- REYNAERT, MATHIAS, AND JAMES M. SALLEE (2021): "Who Benefits When Firms Game Corrective Policies?" *American Economic Journal: Economic Policy*, 13 (1), 372–412. [0961,0996]
- RYAN, CONOR (2020): "How Does Insurance Competition Affect Medical Consumption?" Working paper. [0962]
- SCHENNACH, SUSANNE M. (2016): "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 8 (1), 341–377. [0984]
- SILVER-GREENBERG, JESSICA, AND ROBERT GEBELOFF (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). [0959]
- SMALL, KENNETH A., AND HARVEY S. ROSEN (1981): "Applied Welfare Economics With Discrete Choice Models," *Econometrica*, 49, 105–130. [0975]
- SPENCE, A. MICHAEL (1975): "Monopoly, Quality, and Regulation," *The Bell Journal Of Economics*, 6 (2), 417–429. [0959,0962,0963]
- SWEETING, ANDREW (2009): "The Strategic Timing Incentives of Commercial Radio Stations: An Empirical Analysis Using Multiple Equilibria," *RAND Journal of Economics*, 40 (4), 710–742. [0975]
- TEBALDI, PIETRO, ALEXANDER TORGOVITSKY, AND HANBIN YANG (2023): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," *Econometrica*, 91 (1), 107–146. [0973]
- TOWN, ROBERT, AND SU LIU (2003): "The Welfare Impact of Medicare HMOs," *The RAND Journal of Economics*, 34 (4), 719. [0962,0972]
- TRAIN, KENNETH (2015): "Welfare Calculations in Discrete Choice Models When Anticipated and Experienced Attributes Differ: A Guide With Examples," *Journal of Choice Modelling*, 16, 15–22. [0975]
- VATTER, BENJAMIN (2025): "Supplement to 'Quality Disclosure and Regulation: Scoring Design in Medicare Advantage'," *Econometrica Supplemental Material*, 93, <https://doi.org/10.3982/ECTA21182>. [0965,0966,0986,0997]
- ZAPECHELNYUK, ANDRIY (2020): "Optimal Quality Certification," *American Economic Review: Insights*, 2 (2), 161–176. [0961,0962,0994]

---

Co-editor Kate Ho handled this manuscript.

*Manuscript received 19 September, 2022; final version accepted 6 February, 2025; available online 10 February, 2025.*

*The replication package for this paper is available at <https://doi.org/10.5281/zenodo.14794937>. The authors were granted an exemption to publish parts of their data because either access to these data is restricted or the authors do not have the right to republish them. However, the authors included in the package, on top of the codes and the parts of the data that are not subject to the exemption, a simulated or synthetic dataset that allows running the codes. The Journal checked the data and the codes for their ability to generate all tables and figures in the paper and approved online appendices. Whenever the available data allowed, the Journal also checked for their ability to reproduce the results. However, the synthetic/simulated data are not designed to produce the same results.*