Journal of Economic Behavior and Organization 185 (2021) 234–249

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

What do consumers learn from regulator ratings? Evidence from restaurant hygiene quality disclosures

Tami Kim[a], Daniel Martin[b,*]

[a] University of Virginia Darden School of Business, 100 Darden Blvd., Charlottesville, VA 22901, United States
[b] Northwestern University Kellogg School of Management, 2211 Campus Dr., Evanston IL 60208, United States

ARTICLE INFO

Article history:
Received 14 March 2019
Revised 21 January 2021
Accepted 23 February 2021

Keywords:
Ratings
Disclosure
Certification

ABSTRACT

Regulators use ratings to inform consumers about firms and products across a range of industries, but little is known about what consumers learn from such ratings. We propose and directly study two pathways through which consumers form beliefs about products and firms based on these ratings: inference about the absolute implications of a rating (e.g., does a good rating mean a firm is of high quality?) and inference about the relative implications of a rating (e.g., do few firms have a rating this good?). In the context of restaurant hygiene ratings, we find that consumers form incorrect beliefs along both pathways and that their mis-calibrated beliefs are strongly related to their willingness to pay for a restaurant meal. We also find that their misperceptions can be partially reduced with informational interventions that impact their willingness to pay as well.

1. Introduction

Across a wide range of industries, ratings are used by regulatory bodies to summarize information about products and firms. These ratings are a form of "quality disclosure," which is "an effort by a certification agency to systematically measure and report product quality for a nontrivial percentage of products in a market" (Dranove and Jin 2010). For example, ratings on bonds convey information about the riskiness of such investments, and ratings for nonprofits—such as those provided by Charity Navigator and Impact Matters—convey information about the effectiveness of such organizations. The Centers for Medicare and Medicaid Services (CMS) generate ratings about not only Medicare plans but also health care centers, including hospitals, nursing homes, and dialysis centers. The Environmental Protection Agency uses a 10-point scale to rate a vehicle model's fuel economy and air emissions.

Disclosing such ratings to consumers has been shown to meaningfully impact their decisions. For example, Cutler et al. (2004) found that being identified as a high-mortality hospital through the release of a cardiac surgery reporting system was associated with a decline in the number of patients.[1] Similarly, Perraillon et al. (2019) found that an improvement in ratings of nursing home quality increased new admissions. In an analysis of Medicare advantage contract rating systems, Reid et al. (2013) and Darden and McCarthy (2015) documented a strong relationship between ratings and enrollment decisions.

* Corresponding author.
E-mail address: d-martin@kellogg.northwestern.edu (D. Martin).

[1] Because of this, Dranove et al. (2003) found that quality report cards for coronary artery bypass grafts can lead hospitals to be more strategic about the patients they accept.

https://doi.org/10.1016/j.jebo.2021.02.021
0167-2681/

But little is known about what consumers learn about firms and products from these ratings. We propose and directly study two pathways through which consumers form beliefs about products and firms based on regulatory ratings: inference about the absolute implications of a rating (e.g., does a good rating mean a firm is of high quality?) and inference about the relative implications of a rating (e.g., do few firms have a rating this good?).

The notion that consumers would consider the absolute implications of a rating for the quality of a product or firm is entirely standard. After all, ratings are often meant to be a standalone communication tool that conveys the quality of a product or firm even in the absence of competitor products or firms. It is also natural for consumers to consider the relative implications of a rating if they are searching among firms or if they care about relative quality independently of search, as suggested by a growing literature on rankings. For example, Pope (2009) examines hospital rankings that are released by *U.S. News and World Report* and finds that even when controlling for the absolute score, relative ranking matters for attracting more patients. In a different context, *U.S. News and World Report* college rankings have also been shown to impact application decisions; for instance, a one-rank improvement led to a 1% increase in the number of applications to that college (Luca and Smith, 2015).

While it seems likely that consumers will form correct beliefs from ratings when repeated and clear feedback is provided about the characteristics of a product or firm being rated, such feedback is not always available. Feedback is not repeated when products are infrequently purchased, and feedback is not clear when products are credence goods (e.g., charitable giving, accountants, legal services, vitamin supplements, car repairs, child day care, etc.). In such situations, consumers might struggle to perform correct inference from ratings.

One such case is restaurant hygiene ratings. Almost all state and local health departments provide hygiene ratings of restaurants, and many municipalities now mandate that restaurants display these hygiene ratings prominently at their entrance (Jin and Leslie 2003). This means that hygiene ratings are salient, require little attention or effort to learn, and can be viewed before a customer enters a restaurant. In additional to physical displays of these ratings, restaurant hygiene ratings are now available on some digital platforms. For instance, Yelp.com recently introduced a feature that allows consumers to see the history of hygiene ratings for each restaurant,[2] which can significantly reduce purchase intentions for restaurants with low ratings (Dai and Luca 2020).

We investigate whether consumers form correct beliefs from ratings in this setting by examining whether consumers in California make correct inferences about restaurant hygiene based on San Francisco (SF) Department of Public Health ratings, nearly all of which fall between 71 and 100.[3] We identify failures in understanding the absolute implications of these ratings by having participants guess the fraction of restaurants that had a high risk health violation for a particular range of health ratings. We identify failures in understanding the relative implications of these ratings by having participants guess the distribution of ratings across restaurants, which allows us to infer their beliefs about the fraction of restaurants that have higher health risks than a restaurant with a particular rating.

We find that consumers misperceive both the absolute and relative implications of these ratings. They guess that health risks are much lower than they actually are for low (ratings between 71 and 85) and higher than they actually are for the highest ratings (ratings between 96 and 100). Also, consumers appear to believe that restaurants are relatively better in comparison with other restaurants in terms of health risks than they actually are, with the biggest gap at middle ratings (ratings between 86 and 90). This is likely to matter for market outcomes, as we find evidence of a relationship between their beliefs about absolute and relative health risks and their willingness to pay (WTP) for a meal at the restaurant where they most frequently eat.

We also test the effectiveness of two informational interventions. One aims to reduce misperceptions of the relative implications of these ratings, and it does so partially, with the biggest decrease in misperceptions of relative health risks at middle ratings. Our second intervention aims to reduce both misperceptions, and it achieves a nearly identical reduction in misperceptions about relative health risks as the first intervention. The main impact from this intervention on beliefs about absolute health risks is to increase perceptions of health risks for all ratings. This serves to decrease misperceptions of absolute health risks at low and middle ratings (ratings between 71 and 90), but increase misperceptions of absolute health risks at the highest ratings (ratings between 96 and 100).

What impact do these interventions have on WTP? Both interventions increase the gain in WTP from an increase in ratings and increase the loss in WTP for a decrease in ratings, which provides incentives for firms to improve ratings. For the second intervention, the increase in the loss in WTP for large rating decreases (ratings that decrease by over 20 points) is greater, which means that firms have even stronger incentives not to let ratings fall. Changes in WTP also have consequences for the welfare of consumers. On average, we estimate that WTP would be $2.87 lower in the control treatment if beliefs were correctly calibrated (this differs by score range). The two informational interventions reduce this gap by $0.81 and $0.87, respectively.[4]

[2] See: https://www.forbes.com/sites/andriacheng/2018/07/24/yelp-to-post-health-inspection-scores-on-restaurant-pages-nationwide/.

[3] Restaurants in San Francisco are required to post a sign that shows the restaurant hygiene rating from their most recent San Francisco Department of Public Health inspection.

[4] This estimate is obtained by multiplying the mistakes for each score range reported in Table 2A and Table 3B by the change in WTP estimated in Specification 4 of Table 6, and then weighting the change in WTP for each score range by the proportion of scores in that range, which is given in the first column of Table 3A. This analysis was suggested by a referee, so is not a part of the pre-analysis plan.

To the best of our knowledge, this is the first paper to identify mis-calibrated beliefs about products and firms based on regulatory ratings and the first to investigate the sources of these misperceptions. While we focus on one specific context—restaurant health ratings—our findings can also provide insights about the inferences that consumers make in other quality disclosure settings and for other types of ratings. For example, if consumers incorrectly infer that nursing home quality does not vary much with CMS ratings or believe that Yelp ratings provide a stronger signal of quality than they actually do, this could explain why Yelp ratings have been found to have a bigger impact on resident admissions at nursing homes than CMS ratings (Li et al., 2020).[5] We discuss further the potential ramifications of our results for the design of ratings systems and for educating consumers about these ratings in the Discussion and Conclusion section.

As an additional contribution to the literature, we provide evidence that regulatory ratings influence consumers' WTP by impacting their beliefs about the quality of products and firms. Although this is often assumed to be true, several alternative reasons have been given for why ratings might impact consumer behavior. For example, it has been proposed that variation in WTP with ratings might be due to variation in quality, which consumers observe separately from ratings. Restaurants with higher ratings may get more consumers because of some observable factor related to hygiene, not because they have higher ratings.[6] It has also been proposed that variation in behavior due to ratings can be driven by structural factors. For instance, Perraillon et al. (2019) propose that the lack of a behavioral effect in response to ratings at the low end of ratings in the setting they study could be due to local supply constraints in areas with lower ratings.

2. Model and experiment

2.1. Model of decision-making and predictions

We propose a simple model of decision-making in which a consumer's utility for a product is a quasi-linear function of absolute quality, relative quality, and price:

$$U(absolute\ quality,\ relative\ quality) - p$$

Based on evidence from the literature, we assume that utility is increasing in both absolute quality and relative quality. We also assume that consumers do not directly observe absolute and relative quality, but must infer it from ratings, and that these beliefs are correct. The function that maps ratings to beliefs about absolute quality is A, and the function that maps ratings to beliefs about relative quality is R. For simplicity, we assume there is no uncertainty in beliefs, but this model could easily be amended to incorporate uncertainty.

For a given rating r and outside option with value v, the consumer purchases the good if

$$U(A(r), R(r)) - p > v$$

Thus, WTP is the price at which the consumer is indifferent between purchasing the product and taking the outside option.

We use our experiment to evaluate two main predictions for this model:

1. **Inferences are correct**: The functions A and R correspond to the true map between ratings and absolute and relative quality.
2. **Inferences matter for WTP:** Increases in rating r, $A(r)$, or $R(r)$ correspond with increases in WTP.

2.2. Context

We investigate consumer inference about absolute and relative health risks from the ratings that restaurants received during unscheduled routine inspections in SF, California. The Department of Public Health in SF uses a numeric scoring system in which health scores can range between 0 and 100. During an inspection, a restaurant is evaluated on the number and severity of their health code violations. Violations are divided into three types: low risk, moderate risk, and high risk.

For our pool of inspections, we used all routine unscheduled inspections conducted by the SF Department of Public Health from January 1, 2018, to April 30, 2018, which we retrieved from the publicly available database of inspections on the SF Department of Public Health website (https://www.sfdph.org/). We downloaded this data on May 10, 2018. We chose a four-month range with the aim of capturing sufficient variation in inspection outcomes and to ensure that no single inspection or restaurant was over-represented in our analysis. After removing four inspections that did not have a health score listed, our pool included 1770 routine, unscheduled inspections of 1759 restaurants.

The minimum health score in this pool was 61, while the maximum health score was 100. See Fig. 1 for the full distribution of health scores. The distribution is left-skewed, and small spikes appeared near health scores of 80 and 90.

[5] Li, Lu, and Lu (2020) also find that the relative impact of Yelp ratings and CMS ratings varies across communities depending on Yelp penetration and education levels, both of which could plausibly impact inferences about ratings.

[6] To help distinguish between these forces, Jin and Sorensen (2006) compare publicly released and non-publicly released ratings of health care plans by the National Committee for Quality Assurance (NCQA), a nonprofit firm that provides ratings on health care quality, to show that these ratings impact only a small fraction of enrollment decisions, but that the implied utility gains were high for the individuals. Another possibility is to estimate the change before and after the ratings are released (Chernew et al., 2008; Dranove and Sfekas, 2008; Werner et al., 2016).

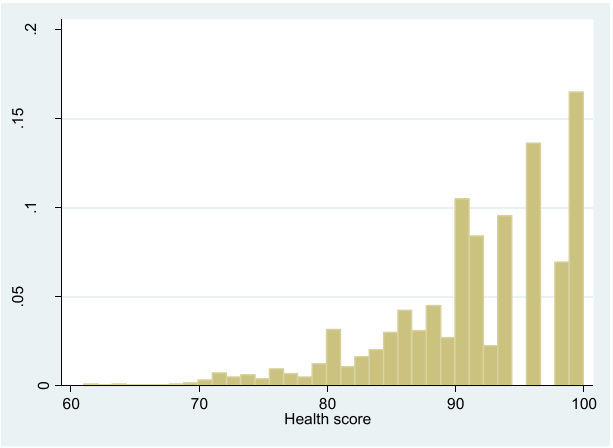

Fig. 1. Distribution of health scores for actual unscheduled, routine restaurant hygiene inspections in San Francisco, California, between January and April 2018.

Table 1
Summary statistics for observable demographics of participants by condition.

VARIABLES	Control			D condition			DD condition		
	N	Mean	Std. dev.	N	Mean	Std. dev.	N	Mean	Std. dev.
Age	379	37.01	13.47	386	35.98	12.36	384	36.33	12.39
Male (dummy)	379	0.441	0.497	386	0.526	0.500	384	0.477	0.500
Within SF area (dummy)	379	0.245	0.431	386	0.241	0.428	384	0.232	0.423
Within LA area (dummy)	379	0.472	0.500	386	0.487	0.500	384	0.466	0.500
>1 restaurant meal/month (dummy)	379	0.689	0.464	386	0.720	0.449	384	0.706	0.456
SF meal in 6 months (dummy)	379	0.401	0.491	386	0.440	0.497	384	0.438	0.497
WTP at most frequent restaurant	379	20.57	14.13	386	21.02	14.27	384	21.88	14.27
Guess of health score at most frequent restaurant	379	86.67	13.04	386	87.01	14.76	384	86.94	12.08

2.3. Method

2.3.1. Participants

1149 participants (52% Female; $M_{age} = 36.4$, $SD_{age} = 12.7$) are in our analysis sample (see Table 1 for summary statistics of participant characteristics). These participants were all part of the Kellogg School of Management panel on Amazon's Mechanical Turk (Kellogg MTurk Panel). Participants were asked for their state of residence before starting the experiment, and only participants who indicated they were California residents were allowed to begin the experiment.

Data collection took place in January 2020. Participants were paid $0.50 as a baseline payment, which is standard for MTurk studies of this duration (the average time to complete the experiment was expected to be under 5 min). Participants could also earn a bonus payment of up to $2, which we describe in more detail below.

2.3.2. Design and procedure

We follow Cavallo et al. (2017) in using many of the best practices identified in the literature for getting high-quality responses when running studies on MTurk. For instance, we used neutral language in our recruitment material, calling this only a "Restaurant Survey" so as not to indicate the purpose of our study. In addition, after consenting to participate in our study, only those participants who answered two attention-check questions correctly were able to proceed with the study. After passing the attention checks, participants were randomly assigned to one of three conditions: "Control"; "Distribution" (D); and "Distribution and Description" (DD).

All participants began by reading the following description: "Every restaurant in San Francisco (California) gets regularly inspected by the Health Department. During each inspection, they receive a Food Safety score, also known as a 'health score.' This score is calculated based on the type and the number of health code violations observed. 100 is the best possible score, and 0 is the worst possible score. During today's session, you will be asked to make guesses about restaurants with different health scores and will receive a bonus payment if your guesses are accurate." Those in the "Distribution" (D) condition were further given a table that showed the percentage of routine, unscheduled restaurant inspections in SF from January to April 2018 that fell in the following four different score ranges: 0–70, 71–85, 86–90, and 91–100 (e.g., 59% of SF inspections receiving a health score above 90). Those in the "Distribution and Description" (DD) condition were given a summary table that included both the percentage of SF inspections for the four different score ranges as well as how to interpret a score that belongs in a certain range (see the Appendix for screenshots of the experimental interface and instructions). We then

asked participants how much they typically pay per person at the restaurant they most frequently go to and what health score they thought the restaurant would receive if inspected.[7]

The rest of the study mainly consisted of three tasks: elicitation of willingness to pay based on scores, the Score Distribution task, and the Health Risk task. These tasks were presented to participants in a random order. After completing these tasks, participants reported their market participation and demographics.

Task 1: Willingness to Pay Based on Scores. Participants faced six scenarios where they were asked their WTP per person at the restaurant they most often visit if that restaurant had a certain set of health scores. Participants were asked (one scenario at a time) to imagine that this restaurant (i.e., the one they most frequently go to) received a health score between: (1) 96 and 100, (2) 91 and 95, (3) 86 and 90, (4) 81 and 85, (5) 76 and 80, and (6) 71 and 75. These six score ranges were presented in a random order. We did not elicit participants' willingness to pay for a health score below the score of 71 because over 99% of the restaurants who had a routine unscheduled inspection received a score of 71 or higher. Participants were then asked to indicate their willingness to pay per person at that restaurant.[8]

Task 2: Score Distribution Task. We measured the perceived distribution of health scores by asking participants, "If we randomly selected 100 restaurants that had unscheduled routine health inspections in San Francisco, how many restaurants do you think would be within each of the following health score ranges?" Participants allocated 100 points across seven ranges: scores in the range (1) 96 to 100, (2) 91 to 95, (3) 86 to 90, (4) 81 to 85, (5) 76 to 80, (6) 71 to 75, and (7) 0 to 70. Participants were further informed that we will randomly select one of the ranges and they will receive a $1 bonus if they guessed within 5 points of the actual number for that range.

Task 3: Health Risk Task. Participants first read: "As part of today's session, you will be reading about 6 different groups of restaurants. For each group of restaurants, we will ask you to make a prediction. We will randomly select one of your six predictions and you will receive a $1 bonus if you guessed within 5 points of the actual number." Participants were then presented with six different scenarios (one scenario at a time). Each scenario asked the same question about a different score range: "Health code violations are classified into three risk categories: 'High risk,' 'Moderate risk,' and 'Low risk.' Suppose that we randomly selected 100 restaurants that had an unscheduled routine health inspection in San Francisco and had received a health score between [*(1) 96 and 100, (2) 91 and 95, (3) 86 and 90, (4) 81 and 85, (5) 76 and 80, or (6) 71 and 75]* (health scores range from 0 to 100). How many of those 100 restaurants do you think would have a 'High risk' violation?" Participants were reminded once again that one of their six predictions would be randomly selected and that they would receive a $1 bonus if they guessed within 5 points of the actual number.

Market Participation. In addition, we asked participants, "In the past six months, how often have you eaten at restaurants?" and "In the past six months, how often have you eaten at restaurants in San Francisco?" The possible responses were "Never," "At least once," "Once a month," "2–3 times a month," and "Once a week or more."

2.3.3. Design considerations

The goal of Task 1 (i.e., Willingness to Pay Based on Scores) was to determine the relationship between WTP and health scores. We decided to fix a reference point (the participant's most frequently visited restaurant) so that as we varied the imagined health scores, WTP would not change for reasons besides the variation in those scores. We used narrow (5-point) ranges of health scores in this task in order to more precisely identify how much the WTP for the reference restaurant changes for a given change in health score. To reduce demand effects, we chose to present each health score range on separate page, and the six scenarios were presented in a random order.

To test the first prediction of our model that participants' beliefs about relative hygiene quality are well-calibrated, we need to estimate the *R* function, which maps ratings to beliefs about relative quality. Thus, the goal of Task 2 (i.e., Score Distribution task) was to determine the relationship between beliefs about relative hygiene quality and health scores for each participant. We could have more directly assessed this by asking participants to indicate the percentage of restaurants with a health score at or above a certain health score, but we did not do so because we deemed that the distribution of scores was more natural for participants to report. As in Task 1, we used 5-point ranges of health scores so that we could match each health score range used in Task 1 with the perceived relative quality for that range in Task 2. Unlike in Task 1, where we presented the six scenarios in a random order on separate pages, it was necessary to present this task on a single page to allow participants to allocate 100 percentage points across the different score ranges.

We also note that the informational intervention provided in the D condition is designed to reduce misunderstandings of relative quality (i.e., to make the *R* function better calibrated to reality). To mimic a likely policy intervention, we used the coarse categorization of scores that the SF Department of Public Health provides on scorecards and their webpage.

To test the first prediction of our model, we also need to estimate the *A* function, which maps ratings to beliefs about absolute quality. Thus, the goal of Task 3 (i.e., Health Risk task) was to determine the relationship between beliefs about absolute hygiene quality and health scores. Again, we used 5-point ranges of health scores to estimate the *A* function for the health score ranges used in the previous task. As in Task 1, we chose to present each health score range on a separate page and present the six scenarios in a random order to reduce demand effects.

[7] Before conducting the analysis, we decided to drop all participants who reported spending $100 per person or more because these entries potentially represent data-entry errors. This excluded 4.5% of the respondents from our sample.

[8] We also decided in advance to drop all responses that are more than 10 times what the participant typically pays because these entries potentially represent data-entry errors. This led us to drop 0.7% of entries.

Table 2
Summary statistics for guesses in the Health Risk task.

Panel A. Actual and guessed percent of restaurants with high risk violations for each health score range (by condition)

VARIABLES	Actual	Control			D condition			DD condition		
	Mean	N	Mean	Std. dev.	N	Mean	Std. dev.	N	Mean	Std. dev.
% w/ high risk in 96–100	0	379	13.06	24.29	386	13.90	24.69	384	15.52	24.90
% w/ high risk in 91–95	17.83	379	13.97	23.00	386	15.48	23.91	384	16.74	23.96
% w/ high risk in 86–90	26.99	379	18.46	23.16	386	18.22	22.53	384	22.54	24.72
% w/ high risk in 81–85	75.88	379	21.52	22.24	386	20.62	22.91	384	25.45	25.04
% w/ high risk in 76–80	79.52	379	27.34	23.29	386	25.49	24.50	384	29.66	25.51
% w/ high risk in 71–75	90.91	379	32.94	25.16	386	28.76	25.40	384	35.40	27.80

Panel B. First guess and later guess of percent with high risk violations (all conditions)

VARIABLES	Actual	Control (first guess)			Control (later guess)		
	Mean	N	Mean	Std. dev.	N	Mean	Std. dev.
% w/ high risk in 96–100	0	193	17.34	23.77	956	13.52	24.76
% w/ high risk in 91–95	17.83	189	17.71	24.98	960	14.95	23.35
% w/ high risk in 86–90	26.99	189	19.73	23.81	960	19.75	23.52
% w/ high risk in 81–85	75.88	186	27.75	27.14	963	21.52	22.61
% w/ high risk in 76–80	79.52	194	29.16	25.43	955	27.16	24.30
% w/ high risk in 71–75	90.91	198	38.03	28.36	951	31.18	25.68

Panel C. Guess of percent with high risk violations if first task or later task (all conditions)

VARIABLES	Actual	Control (first task)			Control (later task)		
	Mean	N	Mean	Std. dev.	N	Mean	Std. dev.
% w/ high risk in 96–100	0	369	14.31	25.51	780	14.09	24.22
% w/ high risk in 91–95	17.83	369	15.99	25.40	780	15.13	22.77
% w/ high risk in 86–90	26.99	369	19.95	25.31	780	19.65	22.69
% w/ high risk in 81–85	75.88	369	22.75	25.03	780	22.43	22.76
% w/ high risk in 76–80	79.52	369	27.12	25.63	780	27.67	23.95
% w/ high risk in 71–75	90.91	369	31.95	25.76	780	32.56	26.52

Further, the informational intervention provided in the DD condition was designed to both reduce misunderstandings of relative quality (to make the *R* function better calibrated to reality) and reduce misunderstandings of absolute quality (to make the *A* function better calibrated to reality). Once again, to mimic a likely policy intervention, we gave the information about each score category available on SF Department of Public Health scorecards and their webpage.

3. Results

3.1. Guesses in the Health Risk task (Control condition)

For participants in the control condition, we first examine inferences about the absolute implications of restaurant hygiene ratings using responses in Task 3 (i.e., the Health Risk task). In this task, participants guess the probability that a restaurant has a high risk health code violation if it received a health score in a particular range (e.g., between 96 and 100). These guesses indicate the extent to which participants believe restaurants present a meaningful health risk based on their health scores.

Panel A of Table 2 shows that the actual probability of high risk health code violations varies widely across health score ranges. No restaurants with a health score between 96 and 100 had a high risk health code violation, but over 90% of restaurants with a health score between 71 and 75 had a high risk health code violation. Like the actual probability, the average guess of the probability decreases monotonically with the health score range.

However, for "low" health score ranges (71–75, 76–80, and 81–85), the average guess is more than 50 percentage points lower than the actual probability. On the other hand, for the highest score range (96–100), the average guess is around 13.5 percentage point higher than the actual probability. These results suggest that participants believe that restaurants with a lower health score are less risky than they actually are and that restaurants with a higher health score are more risky than they actually are. While the average guess is better calibrated for intermediate score ranges, the average guess is statistically different from the actual probability for all score ranges at a 1% confidence level using a two-sided *t*-test. These misperceptions are represented graphically in Fig. 2.

We use regression analysis to confirm that these patterns are robust to controlling for demographic characteristics (either directly or using participant fixed effects). The first two columns of Table 4 provide estimates of participant misperceptions, measured as the difference between the actual rate and guesses in the Health Risk task, for each health score range. The constant in these regressions provides the estimated under-guess of health risk for "middle" health scores (between 86 and 90), and there is strong evidence of under-guessing of health risk for this range when controlling for differences across

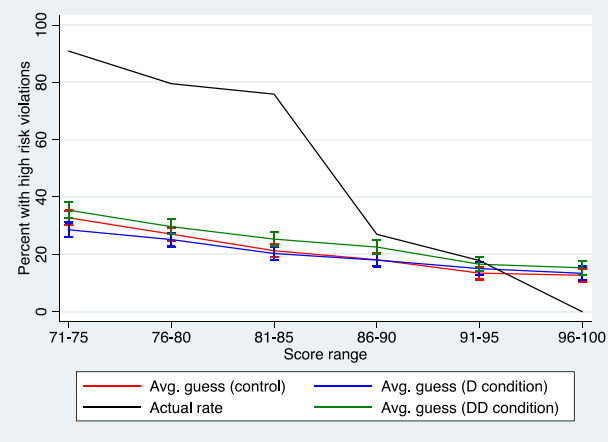

Fig. 2. Actual and guessed percent of restaurants with high risk violations for each health score range (by condition).

participants. Each of the coefficients provides the difference between the under-guess for this range and all of the other health score ranges. Consistent with the averages in Table 2, we find strong evidence that under-guessing increases for lower health score ranges and decreases for higher health score ranges. The decrease is enough for health scores between 96 and 100 for the net effect to produce over-guessing of health risk (8.891–21.517 = −12.626 for the regression with participant fixed effects).

Because the only source of variation across questions in this task was in the health score ranges, there may be concerns of demand effects, even though guesses were incentivized. Specifically, one might conjecture that demand effects would lead to an increase in the variation in guesses across health score ranges. Thus, if there are demand effects of this type occurring, then participants might be even worse calibrated about absolute health risk than our results suggest.

To look for evidence of demand effects, we examine just the first guess made by a participant in this task. For the first guess, the only demand effects would be forward-looking ones, and since we do not tell subjects what health score ranges they will be evaluating in advance, this potentially reduces the likelihood of forward-looking demand effects. Panel B of Table 2 compares the average first guesses with the average of later guesses in this task, and shows that the range of scores is similar between first guesses and later guesses, which provides little evidence that demand effects are producing greater dispersion in later guesses. The only substantive difference appears to be that first guesses are slightly higher for all score ranges.

Also, because participants completed tasks in a random order, there may be concerns about order effects. Panel C of Table 2 compares average guesses in this task if this task was completed first or later, and the averages are quite similar between the two and not statistically different (at a 10% confidence level using a two-sided *t*-test).

3.2. Guesses in the Score Distribution Task (Control condition)

Next, we examine inferences about the relative implications of restaurant hygiene ratings using responses in Task 2 (i.e., the Score Distribution task). In this task, participants guessed the percentage of restaurants in each health score range.

Panel A of Table 3 shows that the average guess of the percentage of restaurants with a health score between 91 and 95 is close to actual rate (21.9% vs. 18.4%), but also that the average guesses of the percentage of restaurants with health scores above 85 is too low and the average guess of the percentage of restaurants with health scores below 86 is too high.

What does this tell us about participant beliefs about the relative implications of restaurant hygiene ratings? For this, we examine what these guesses imply about the percentage of restaurants with a health score at or above a certain number. These percentages indicate how relatively risky participants believe restaurants are based on their health score.

Panel B of Table 3 reports the implied guess of the percentage of restaurants at or above each health score range, and Fig. 3 represents these patterns graphically. Both show that for all of these health score ranges, it is as if participants believe restaurants are relatively better than they actually are. The implied percentage is statistically different from the actual percentage for all score ranges at a 1% confidence level using a two-sided *t*-test.

Because participants completed tasks in a random order, there may be concerns about order effects for this task as well. We find that for all health score ranges, average guesses are not statistically different if the distribution of ratings task occurred first (at a 10% confidence level using a two-sided *t*-test).

Panel B of Table 3 and Fig. 3 also indicate that the gap is larger for middle health score ranges (between 86 and 90). This is confirmed with a regression analysis that controls for demographic characteristics (both directly and using participant fixed effects). The third and fourth columns of Table 4 provide estimates of participant misperceptions, which are measured as the difference between the actual rate and implied guesses of the percentage of restaurants at or above each health score range. The constant provides the estimated under-guess for health scores between 86 and 90, and there is strong evidence of

Table 3
Summary statistics for guesses in Score Distribution task (by condition).

Panel A. Actual and guessed percent of restaurants at health score range

VARIABLES	Actual	Control			D condition			DD condition		
	Mean	N	Mean	Std. dev.	N	Mean	Std. dev.	N	Mean	Std. dev.
% at 96–100	37.06	379	16.61	19.02	386	21.64	18.26	384	20.78	17.86
% at 91–95	21.87	379	18.42	14.03	386	22.58	13.59	384	23.26	14.17
% at 86–90	23.44	379	17.38	10.24	386	17.09	9.803	384	17.64	9.138
% at 81–85	9.610	379	15.11	9.229	386	13.02	7.509	384	13.31	7.932
% at 76–80	4.690	379	12.09	8.565	386	10.07	7.317	384	10.32	7.612
% at 71–75	2.480	379	9.120	7.304	386	7.894	6.990	384	7.737	6.947
% at 0–70	0.850	379	11.27	13.89	386	7.707	10.70	384	6.943	9.369

Panel B. Actual and implied guess of percent at or above each health score range

VARIABLES	Actual	Control			D condition			DD condition		
	Mean	N	Mean	Std. dev.	N	Mean	Std. dev.	N	Mean	Std. dev.
% at 96–100	37.06	379	16.61	19.02	386	21.64	18.26	384	20.78	17.86
% at or above 91–95	58.93	379	35.03	24.28	386	44.22	21.68	384	44.04	22.07
% at or above 86–90	82.37	379	52.41	24.04	386	61.30	21.74	384	61.68	20.94
% at or above 81–85	91.98	379	67.52	21.95	386	74.33	19.65	384	75.00	17.47
% at or above 76–80	96.67	379	79.61	18.54	386	84.40	15.30	384	85.32	13.16
% at or above 71–75	99.15	379	88.73	13.89	386	92.29	10.70	384	93.06	9.369

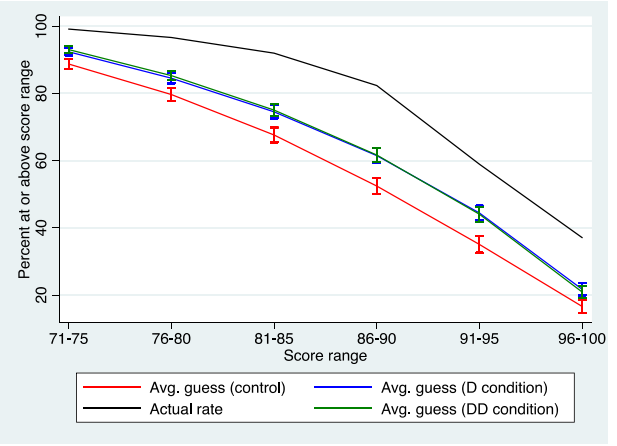

Fig. 3. Actual and implied guess of percent at or above each health score range (by condition).

Table 4
Regressions of participant guesses (control condition).

VARIABLES	(1) Actual guess% w/ high risk	(2) Actual guess% high w/ risk	(3) Actual guess% at or above	(4) Actual guess% at or above
Health score in 71–75	49.287***	49.287***	−19.532***	−19.532***
	(1.257)	(1.375)	(0.927)	(1.014)
Health score in 76–80	43.573***	43.573***	−12.936***	−12.936***
	(0.934)	(1.022)	(0.763)	(0.835)
Health score in 81–85	45.727***	45.726***	−5.549***	−5.536***
	(0.792)	(0.866)	(0.479)	(0.523)
Health score in 91–95	−4.520***	−4.521***	−6.048***	−6.035***
	(0.734)	(0.803)	(0.532)	(0.581)
Health score in 96–100	−21.587***	−21.517***	−9.458***	−9.454***
	(0.728)	(0.793)	(0.937)	(1.020)
Constant	11.845***	8.891***	33.009***	29.910***
	(2.910)	(0.570)	(3.067)	(0.477)
Observations	2258	2258	2258	2258
R-squared	0.620	0.882	0.095	0.774
Demographic controls	Yes	No	Yes	No
Participant fixed effects	No	Yes	No	Yes

Notes: The dependent variable is the under-guess (actual guess) in the task. In parentheses are robust standard errors clustered by participant. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Demographic controls are dummies for age quartile, gender, whether located in San Francisco or Los Angeles, whether had more than one meal in a restaurant per month, and whether had a restaurant meal in San Francisco in the past six months.

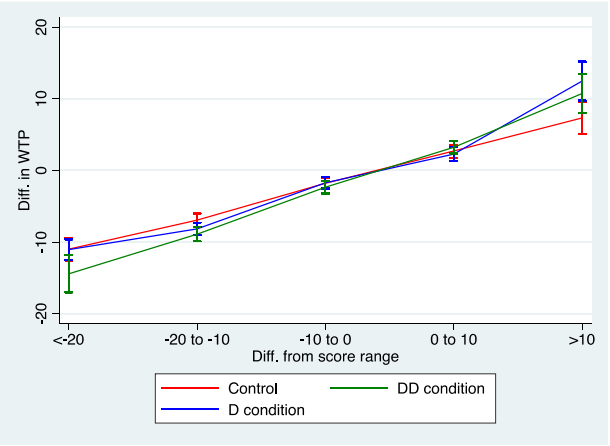

Fig. 4. Difference in WTP for each range of differences in health score (relative to WTP and guess of health score for most frequently attended restaurant).

under-guessing for this range when controlling for differences across participants. Each of the coefficients provides the difference between the under-guess for this range and all of the other health score ranges. We find strong evidence that there is less under-guessing for health score ranges above and below this range. In addition, the point estimates are increasing as the health score ranges move further away, but this is not enough to predict over-guessing on net.

3.3. Willingness to Pay (Control condition)

Next, we examine the elicited WTP of participants in the control condition. As shown in Table 1, participants in this condition reported spending $20.57 per person at the restaurant where they most frequently eat, and on average they guessed that this restaurant would receive a health score of 87. When asked what their WTP would be if they learned this restaurant received a health score in different ranges, there was substantial variation in WTP: for instance, for hygiene ratings between 76 and 80, the average WTP was $15.44, for hygiene ratings between 86 and 100, the average WTP was $20.50, and for hygiene ratings between 96 and 100, the average WTP was $25.54.

We find similar effects in regressions that control for demographic characteristics (both directly and using participant fixed effects). The first two columns of Table 6 show the estimates of WTP at each score range relative to the WTP for middle health scores (between 86 and 90). For the regression with participant fixed effects, the constant shows that our estimate of WTP for health scores between 86 and 90 is $20.48. The regression coefficients also show that increasing the health score range to between 96 and 100 increases WTP by $5.18 and that decreasing the health score range to between 76 and 80 decreases WTP by $5.05.

The third and fourth columns of Table 6 show that beliefs about absolute and relative health risks are strongly related to WTP. However, these estimates are sensitive to which specification is used. The sign on the coefficient for beliefs about absolute health risks differs when we control for demographic characteristics directly (Specification 3) and when we use participant fixed effects (Specification 4). Because participant fixed effects allow us to control for all factors that vary across participants, this is the more appealing specification, and sensibly, when we use participant fixed effects, higher beliefs about absolute health risks correspond with lower WTP.

However, there is a natural concern with these regression specifications. There may be a kink point in the relationship between WTP and health score around the health score of the restaurant where participants most frequently eat, and this could generate non-linearities in the relationship between health score and WTP.[9] To address this possibility, we look at the relationship between the changes in WTP and perceptions of absolute and relative health risk at the most frequently attended restaurant as the health score changes for the scenarios in this task. To infer the perceptions of absolute and relative health risk of the restaurant where participants most frequently eat, we match the participant's guess of the health score of that restaurant with their elicited beliefs about absolute and relative health risk from the other tasks.

Fig. 4 shows how WTP changes at the most frequently attended restaurant if that restaurant's health score changes. On the *y*-axis of Fig. 4, we give the difference between the typical WTP at the most frequently attended restaurant at the WTP for each possible health score range, and on the *x*-axis, we give the difference between the guess of the health score at the most frequently attended restaurant and each possible health score range. For the control condition, this relationship appears roughly linear.

In the first two columns of Table 7, we show the estimates of the differences in WTP for each range of differences in health score (relative a difference in health scores between −10 and 0) based on regressions that control for demographic

[9] We thank an anonymous referee for raising this concern and for suggesting the approach we use to address it.

Table 5
Reported WTP by health score range.

Panel A. Reported WTP for each health score range (by condition)

VARIABLES	Control			D condition			DD condition		
	N	Mean	Std. dev.	N	Mean	Std. dev.	N	Mean	Std. dev.
Health score in 96–100	375	25.54	19.13	381	26.35	20.03	381	27.32	20.53
Health score in 91–95	376	23.76	18.62	381	24.19	18.70	383	25.93	20.60
Health score in 86–90	377	20.50	17.27	382	21.60	19.09	384	21.88	18.92
Health score in 81–85	376	18.30	17.28	381	18.09	17.19	383	18.55	18.11
Health score in 76–80	377	15.44	15.77	381	15.65	17.53	384	15.74	17.13
Health score in 71–75	377	13.22	15.03	380	13.59	17.72	384	13.47	16.87

Panel B. First report and later reports of WTP (all conditions)

VARIABLES	Control (first guess)			Control (later guess)		
	N	Mean	Std. dev.	N	Mean	Std. dev.
Health score in 96–100	205	26.90	19.79	944	27.42	22.81
Health score in 91–95	171	27.25	22.03	978	24.65	19.48
Health score in 86–90	162	19.83	14.55	987	21.90	19.44
Health score in 81–85	196	17.69	16.44	953	19.00	18.62
Health score in 76–80	204	15.50	16.85	945	16.08	17.58
Health score in 71–75	211	14.81	18.54	938	13.67	17.21

Panel C. Report of WTP if first task or later task (all conditions)

VARIABLES	Control (first task)			Control (later task)		
	N	Mean	Std. dev.	N	Mean	Std. dev.
Health score in 96–100	359	27.67	22.36	790	27.17	22.27
Health score in 91–95	359	25.55	20.74	790	24.81	19.50
Health score in 86–90	359	22.36	20.38	790	21.26	18.10
Health score in 81–85	359	19.38	19.36	790	18.51	17.76
Health score in 76–80	359	16.78	18.46	790	15.61	16.96
Health score in 71–75	359	14.59	18.17	790	13.56	17.13

characteristics (both directly and using participant fixed effects). The difference in WTP appears to decrease and increase symmetrically when moving away from the difference in WTP for a score difference between −10 and 0, as suggested by Fig. 4.

However, one noticeable aspect of using this difference-based approach is that participant fixed effects are no longer necessary to produce a sensible sign for absolute health risks. This provides both robustness for the sign of the relationship and also supports the use of a difference-based approach. This will be useful in subsequent specifications where it is not possible to use fixed effects.

As with choices in Task 3 (i.e., the Health Risk task), there may be concerns of demand effects in WTP reports because the only characteristic that varied across questions was health score ranges. To look for evidence of demand effects, we compare the first WTP report made by participants in this task to later WTP reports. For demand effects to impact the very first reported WTP, they would need to be forward looking, and since we do not tell subjects what health score ranges they will be evaluating in advance, this potentially reduces the likelihood of forward-looking demand effects. Panel B of Table 5 compares the average of first WTP reports with the average of later WTP reports in this task, and it shows that reports are similar for all score ranges. For each score range, the average of WTP reports is not statistically different if the risk assessment task occurred first or later (at a 10% confidence level using a two-sided *t*-test). Thus, there is little evidence of greater dispersion across score ranges in later guesses. However, there may be remaining concerns about demand because reports of WTP were not incentivized. To help alleviate these concerns, we ran a follow-up study with a smaller sample in which WTP was incentivized, and we find a similar strong relationship between ratings and WTP in that study as well (full details are given in the Appendix).

Also, because participants completed tasks in a random order, there may be concerns about order effects in this task as well. Panel C of Table 5 compares average guesses in this task if this task was completed first or later, and the averages are quite similar between the two. Not surprisingly, average guesses are not statistically different if the risk assessment task occurred first (at a 10% confidence level using a two-sided *t*-test).

3.4. Interventions (D and DD conditions)

Next, we examine the impact of our two informational interventions. In the "Distribution" (D) condition, participants were shown at the beginning of the experiment a table that showed the percentage of SF inspections for four different score ranges: 0–70, 71–85, 86–90, and 91–100. In the "Distribution and Description" (DD) condition, participants were also

Table 6
Regressions of WTP (control condition).

VARIABLES	(1) WTP	(2) WTP	(3) WTP	(4) WTP
Health score in 71–75	−7.273***	−7.273***		
	(0.408)	(0.447)		
Health score in 76–80	−5.054***	−5.054***		
	(0.374)	(0.409)		
Health score in 81–85	−2.186***	−2.207***		
	(0.318)	(0.344)		
Health score in 91–95	3.267***	3.246***		
	(0.391)	(0.425)		
Health score in 96–100	5.072***	5.183***		
	(0.400)	(0.418)		
Guess of% w/ high risk			0.196***	−0.030**
			(0.050)	(0.015)
Guess of% at or above score range			−0.145***	−0.141***
			(0.017)	(0.009)
Constant	19.455***	20.478***	22.700***	28.101***
	(2.138)	(0.226)	(2.215)	(0.476)
Observations	2258	2258	2258	2258
R-squared	0.204	0.912	0.241	0.903
Demographic controls	Yes	No	Yes	No
Participant fixed effects	No	Yes	No	Yes

Notes: The dependent variable is the WTP at a given score range. In parentheses are robust standard errors clustered by participant. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Demographic controls are dummies for age quartile, gender, whether located in San Francisco or Los Angeles, whether had more than one meal in a restaurant per month, and whether had a restaurant meal in San Francisco in the past six months.

Table 7
Regressions of Difference in WTP (control condition).

VARIABLES	(1) Diff WTP	(2) Diff WTP	(3) Diff WTP	(4) Diff WTP
Difference in health score <20	−8.739***	−9.612***		
	(1.015)	(0.708)		
Difference in health score −20 to −10	−4.959***	−5.450***		
	(0.470)	(0.414)		
Difference in health score 0 to 10	4.376***	4.684***		
	(0.394)	(0.400)		
Difference in health score >10	8.234***	9.402***		
	(2.237)	(0.948)		
Diff in guess of% w/ high risk			−0.038**	−0.038***
			(0.019)	(0.014)
Diff in guess of% at or above			−0.112***	−0.140***
			(0.011)	(0.009)
Constant	−0.132	−1.798***	−0.533	−1.106***
	(1.733)	(0.235)	(1.378)	(0.046)
Observations	2258	2258	2092	2092
R-squared	0.178	0.827	0.178	0.804
Demographic controls	Yes	No	Yes	No
Participant fixed effects	No	Yes	No	Yes

Notes: The dependent variable is the difference in WTP at a given score range from the WTP at the most frequent restaurant. In parentheses are robust standard errors clustered by participant. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Demographic controls are dummies for age quartile, gender, whether located in San Francisco or Los Angeles, whether had more than one meal in a restaurant per month, and whether had a restaurant meal in San Francisco in the past six months.

provided the SF Department of Public Health labels for these ranges (Poor, Needs improvement, Adequate, and Good) and the SF Department of Public Health descriptions of the health risks for each range.

There are several reasons why these interventions might not successfully eliminate misperceptions. First, participants may overlook this information or simply not remember it later in the experiment, since the details were just presented briefly at the beginning of the experiment. Second, misperceptions may be resistant to new information if participants are conservative in updating their beliefs. There are many examples from the field where strong information disclosures produce surprisingly small effects (for example, see Adams et al., 2020). Third, the information in these interventions was coarser than what we asked of participants in the Health Risk task and Score Distribution task.

Table 8
Regressions of participant guesses (all conditions).

VARIABLES	(1) Actual guess% w/ high risk	(2) Actual guess% w/ high risk	(3) Actual guess% at or above	(4) Actual guess% at or above
Health score in 71–75	51.230***	49.287***	−16.072***	−19.532***
	(0.717)	(1.255)	(0.510)	(0.925)
Health score in 76–80	44.802***	43.573***	−10.337***	−12.936***
	(0.563)	(0.932)	(0.396)	(0.762)
Health score in 81–85	46.154***	45.724***	−4.236***	−5.547***
	(0.456)	(0.790)	(0.247)	(0.478)
Health score in 91–95	−4.622***	−4.523***	−6.059***	−6.047***
	(0.483)	(0.733)	(0.291)	(0.531)
Health score in 96–100	−21.236***	−21.571***	−6.475***	−9.447***
	(0.500)	(0.725)	(0.519)	(0.934)
Condition D	0.933	0.206	−6.348***	−8.888***
	(1.379)	(1.589)	(1.225)	(1.655)
Health score in 71–75 *Condition D		4.042**		5.393***
		(1.694)		(1.241)
Health score in 76–80 *Condition D		1.836		4.159***
		(1.328)		(0.978)
Health score in 81–85 *Condition D		0.915		2.102***
		(1.050)		(0.617)
Health score in 91–95 *Condition D		−1.639		−0.312
		(1.127)		(0.734)
Health score in 96–100 *Condition D		−0.792		3.905***
		(1.135)		(1.297)
Condition DD	−3.117**	−4.301**	−6.627***	−9.224***
	(1.436)	(1.699)	(1.188)	(1.626)
Health score in 71–75 *Condition DD		1.772		4.939***
		(1.820)		(1.272)
Health score in 76–80 *Condition DD		1.835		3.601***
		(1.396)		(1.008)
Health score in 81–85 *Condition DD		0.370		1.814***
		(1.180)		(0.627)
Health score in 91–95 *Condition DD		1.335		0.275
		(1.162)		(0.711)
Health score in 96–100 *Condition DD		1.792		4.962***
		(1.213)		(1.259)
Constant	7.603***	8.246***	30.018***	31.741***
	(1.964)	(2.013)	(1.742)	(1.864)
Observations	6843	6843	6843	6843
R-squared	0.604	0.605	0.102	0.104
Demographic controls	Yes	Yes	Yes	Yes
Participant fixed effects	No	No	No	No

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Demographic controls are dummies for age quartile, gender, whether located in San Francisco or Los Angeles, whether had more than one meal in a restaurant per month, and whether had a restaurant meal in San Francisco in the past six months.

As shown in Fig. 2, the two interventions only have a small effect on misperceptions of absolute health risks, but these effects differ. The line for the D condition closely tracks the line for the control condition, with a small increase in misperceptions for the lowest health score range. On the other hand, the line for the DD condition is above both lines for all health scores, which means that it reduces misperceptions for low and middle health scores, but increases misperceptions for the highest health scores.

As shown in Fig. 3, the two interventions have a larger effect on correcting misperceptions of relative health risks, and these effects are nearly identical on average. The largest reduction in misperceptions appears to occur for middle health scores.

We find that these results are statistically significant by running regressions that control for demographic characteristics.[10] As shown in the first column of Table 8, there is not strong evidence of an aggregate effect on beliefs of absolute health risks from providing the distribution of health scores (the D condition), but there is strong evidence of aggregate decrease in the difference between actual and guesses of absolute health risks from also providing the descriptions of these health scores (the DD condition). However, if we look at the effect by health score, we see evidence of nuanced non-linear effects. For the lowest health score range, there is evidence of an increase in the difference between actual and guesses of absolute health risks. That is, there is an increase in the misperceptions of absolute health risks. On the other hand, the

[10] It is not possible to use participant fixed effects for this analysis, as the intervention does not vary within participant.

Table 9
Regressions of Difference in WTP (all conditions).

VARIABLES	(1) Diff WTP	(2) Diff WTP
Difference in health score <20	−9.596***	−8.486***
	(0.676)	(1.053)
Difference in health score −20 to −10	−5.897***	−4.835***
	(0.278)	(0.471)
Difference in health score 0 to 10	4.582***	4.320***
	(0.251)	(0.400)
Difference in health score >10	10.347***	7.913***
	(1.443)	(2.238)
Condition D	0.168	0.149
	(0.816)	(0.720)
Difference in health score <20 *D		−0.595
		(1.440)
Difference −20 to −10 *D		−1.465**
		(0.710)
Difference 0 to 10 *D		−0.508
		(0.583)
Difference >10 *D		3.884
		(3.518)
Condition DD	−0.304	−0.533
	(0.826)	(0.768)
Difference in health score <20 *DD		−3.487*
		(1.968)
Difference −20 to −10 *DD		−1.641**
		(0.664)
Difference 0 to 10 *DD		1.187*
		(0.620)
Difference >10 *DD		3.615
		(3.583)
Constant	1.546	1.683
	(1.160)	(1.174)
	6843	6843
Observations	0.220	0.223
R-squared	No	No
Demographic controls	Yes	Yes
Participant fixed effects	−9.596***	−8.486***

Notes: The dependent variable is the under-guess (actual guess) in the task. In parentheses are robust standard errors clustered by participant. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Demographic controls are dummies for age quartile, gender, whether located in San Francisco or Los Angeles, whether had more than one meal in a restaurant per month, and whether had a restaurant meal in San Francisco in the past six months.

aggregate decrease in the difference between actual and guesses is not significantly different between any of health score ranges and health scores between 86 and 90. However, the estimated decrease is largest for health scores between 81 and 90. This means that there is a decrease in misperceptions for low and middle health scores (e.g., 71–90), but an increase for high health scores (96–100).

Looking at the third column of Table 8, we see there is strong evidence of an aggregate decrease in the difference between actual and guesses of relative health risks for both interventions. The effects are also of similar size. We also see similar effects when considering the impact of these interventions for each health score range. These strongest decreases are for health scores between 86 and 95, and the decreases in misperceptions are significantly smaller for other health score ranges. However, the net effect of the intervention is still negative for every health score range, which mean that there is a decrease in misperceptions for all of these health score ranges.

Finally, we look at the impact of these interventions on WTP. Because we cannot run participant fixed effects when considering treatment effects, we use the difference-based approach to study WTP. As shown in Fig. 4, the D condition largely tracks the control condition, but there is a larger increase in WTP at the highest increase in health scores. On the other hand, the DD condition both decreases WTP for the largest decrease in health scores and increases WTP for the highest increase in health scores.

Table 9 shows the effect of the interventions on WTP, controlling for participant demographics. Specifically, there is not strong evidence of an aggregate effect of either intervention at the aggregate level, but by looking at the effect of each difference in health score, we see a more nuanced picture. For the D condition, there is a large but statistically insignificant impact for the largest increases in health score. There is evidence of a drop for decreases in health score, but not for the

largest decreases in health scores. For the DD condition, there is again a large but statistically insignificant impact for the largest increases in health score. However, we now see evidence of a drop in WTP for decreases in health score, including the largest decreases in health scores. All of these effects are relative to the impact of the intervention for a drop in health scores between 0 and 10, which is not statistically significant in either treatment.

3.5. Attentional robustness checks

Finally, we performed a robustness check where we exclude all participants that exhibited any of three patterns that are potentially revealing of low attentional effort: non-changing responses in the Health Risk task, non-monotonic WTP with respect to increases in health score, and non-monotonic responses in the Health Risk task with respect to increases in health score. We find that our primary regression results are robust to excluding these participants (see Tables A2-A4 and the corresponding analysis provided in the Appendix).[11]

It should be noted that while non-changing responses and monotonicity violations could reflect attentional limitations, they may also represent non-monotonicity in beliefs/preferences or general stochasticity in choice. There is growing evidence of stochasticity in choice in standard lab settings (e.g., Agranov and Ortoleva 2017). Monotonicity violations have been observed in other experiments in which otherwise ordered questions have been presented one at a time and in a scrambled order. For example, two recent lab experiments presented the Holt-Laury risk assessment task, which is traditionally presented as an ordered list of binary lottery choices, in a scrambled order as a list (Freeman and Mayraz 2019), and in a scrambled order over several screens (Brown and Healy 2018). In the latter case, the authors found a substantial number of monotonicity violations (32.8%), even with a participant pool and remuneration that are standard for lab experiments.

4. Discussion and conclusion

Across a range of industries, consumers encounter firms and products that have been rated by a regulator. In a world where firms and products only differ in terms of ratings, there is no need for consumers to think about what these ratings imply about firms and products because a simple comparison of two ratings is all that is required for any decision. However, consumers often face tradeoffs—for instance, between ratings and price—which means they need to form correct beliefs from ratings in order to choose the best option for them.

We provide a simple, rational model of decision-making in which (1) a consumer's utility is impacted by absolute and relative quality and (2) importantly, beliefs about absolute and relative quality are inferred correctly from ratings. Using an online experiment in the context of restaurant hygiene ratings, we document an important departure from this model: consumers make systematically incorrect inferences about absolute and relative quality from these ratings. To the best of our knowledge, what consumers learn from regulatory ratings has not been directly studied before, so this is the first documented case of incorrect inference from regulatory ratings in the literature. For this ratings system, we find that consumers think firms with low ratings (85 and below) are higher quality than they are in truth, and firms with the highest ratings (96 and above) are lower quality than they are in truth. Moreover, consumers think firms are relatively better than they are in truth, which is reflected in a belief that the distribution of ratings is not as skewed at it is actually.

Such mis-calibration can lead to consumers having a WTP for a product that is not appropriate given the absolute and relative quality of that product. It can also have substantial implications for markets. Incorrect inference about ratings could influence whether ratings are reported by firms when disclosure is voluntary.[13] Also, inference about ratings could impact how extensively consumers search for products because it determines whether they think they have found a product (or restaurant) that is "good enough."

Our findings could also have substantial ramifications for the design of regulatory rating systems.[12] If beliefs about the distribution of ratings are heterogeneous and mis-calibrated, as they are in the setting we study, then the designers of ratings systems must care not only about the actual distribution of ratings, which they determine by setting the requirements for a particular rating, but also perceptions of the distribution of ratings. As we show, it may be possible to improve inference about ratings by disclosing the distribution of ratings to consumers. An alternative, more indirect, way to close this gap is to disclose (or require firms to disclose) information about the requirements or algorithm that determines the distribution of ratings. However, presenting this information can be complex and may lead consumers to make poor inference (Jin et al., 2018).

Our results also suggest when and why regulators might feel pressure from the entities they regulate to skew ratings. If consumers have correct beliefs about the distribution of ratings, then firms with middling quality can benefit from coarse and inflated ratings systems because they are pooled with better firms that have high ratings. For example, if even middling-quality firms receive a score of 98, and all scores between 98 and 100 get an "A" rating, then middling-quality firms get to share the same rating with the very best firms. However, if consumers have incorrect beliefs about the distribution of ratings, then even when ratings are not coarse, firms with middling quality could benefit from an inflated ratings system. For

[11] This analysis was a suggested by a referee, so is not a part of the pre-analysis plan.

[13] This could potentially interact with incorrect strategic inferrences about non-disclosure that have been documented in lab experiments (e.g., Jin et al., 2015).

[12] See Tadelis (2016) for a discussion of design issues for online review systems.

example, if consumers consistently believe that few firms get a score above 98, perhaps due to experience with educational grading systems, then middling-quality firms might push to receive such high scores, even if this would make no difference in a world with correctly calibrated consumers. Tadelis (2016) articulates this challenge when he writes, "Naively, one may think that a score of 98% [on eBay] is excellent." As Nosko and Tadelis (2015) show, a seller with a score of 98% on eBay is actually in the bottom 10% of sellers. Thus, if a buyer with mis-calibrated beliefs about the distribution of seller scores encounters a seller with a score of 90%, the buyer may (incorrectly) infer that the seller is better than many others, which greatly benefits such sellers. Thus, if such mis-calibration holds more generally, firms with lower ratings might push for inflated ratings. This provides another reason on top of equilibrium forces why "grade inflation" might occur (Li, 2007).

While the two pathways we identified will likely generalize across contexts, future research is needed to better understand how the importance of each pathway might depend on the context. For instance, absolute implications are likely to matter more when products are complex, and relative implications are likely to matter more when ratings are skewed. Future research should also examine whether the two pathways work similarly across different type of goods. For instance, do the absolute and relative pathways also matter in contexts where consumers have a lot of experience with the focal product (or similar products)?

Furthermore, future research should identify factors that produce erroneous inference along the absolute and relative pathways. For instance, there are many rating systems that depend on user-generated ratings of subjective experiences. Inference in these settings brings extra challenges. It is possible that factors such as selection of which consumers purchase a product and who leaves ratings can impact the absolute pathways, while factors such as how ratings are aggregated could impact the role of relative pathways.

Declaration of Competing Interest

None.

Acknowledgments

We would like to give a special thanks to Michael Luca for his thoughtful insights during the early stages of this project, to Kellogg Research Support (particularly Mac Abruzzo, Kat Baker, Ginger Jascobson, and Sam Smith) and Rhys Aglio for their valuable assistance during this project, and to Ginger Zhe Jin, Andrés Espitia De la Hoz, and Román Antonio Acosta for helpful comments on the draft. Our study was approved by the IRB at Northwestern University and preregistered at aspredicted.org (#34266).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2021.02.021.

References

Adams, P., Huntz, S., Palmer, S., Zaliauskas, R., 2020. Testing the effectiveness of consumer financial disclosure: experimental evidence from savings accounts. J. Financ. Econ. In press.
Agranov, M., Ortoleva, P., 2017. Stochastic choice and preferences for randomization. J. Polit. Econ. 125 (1), 40–68.
Brown, A.L., Healy, P.J., 2018. Separated decisions. Eur. Econ. Rev. 101, 20–34.
Cavallo, A., Cruces, G., Perez-Truglia, R., 2017. Inflation expectations, learning, and supermarket prices: evidence from survey experiments. Am. Econ. J. Macroecon. 9 (3), 1–35.
Chernew, M., Gowrisankaran, G., Scanlon, D.P., 2008. Learning and the value of information: evidence from health plan report cards. J. Econom 144 (1), 156–174.
Cutler, D.M., Huckman, R.S., Landrum, M.B., 2004. The role of information in medical markets: an analysis of publicly reported outcomes in cardiac surgery. Am. Econ. Rev. 94 (2), 342–346.
Dai, W., Luca, M., 2020. Digitizing disclosure: the case of restaurant hygiene scores. Am. Econ. J. Microecon. 12 (2), 41–59.
Darden, M., McCarthy, I.M., 2015. The star treatment estimating the impact of star ratings on Medicare Advantage enrollments. J. Hum. Resour. 50 (4), 980–1008.
Dranove, D., Jin, G.Z., 2010. Quality disclosure and certification: theory and practice. J. Econ. Perspect. 48 (4), 935–963.
Dranove, D., Kessler, D., McClellan, M., Satterthwaite, M., 2003. Is more information better? The effects of "report cards" on health care providers. J. Polit. Econ. 111 (3), 555–588.
Dranove, D., Sfekas, A., 2008. Start spreading the news: a structural estimate of the effects of New York hospital report cards. J. Health Econ. 27 (5), 1201–1207.
Freeman, D.J., Mayraz, G., 2019. Why choice lists increase risk taking. Exp. Econ. 22 (1), 131–154.
Jin, G.Z., Leslie, P., 2003. The effect of information on product quality: evidence from restaurant hygiene grade cards. Q. J. Econ. 118 (2), 409–451.
Jin G.Z., Luca M., Martin D., 2015. Is No News (Perceived as) Bad News? An Experimental Investigation of Information Disclosure. Am. Econ. J. Microecon. NBER Working Paper No. 21099. In press.
Jin G.Z., Luca M., Martin D., 2018. Complex Disclosure. Manage. Sci. NBER Working Paper No. 24675. In press.
Jin, G.Z., Sorensen, A.T., 2006. Information and consumer choice: the value of publicized health plan ratings. J. Health Econ. 25 (2), 248–275.
Li, H., 2007. A signaling theory of grade inflation. Int. Econ. Rev. 48 (3), 1065–1090.
Li, Y., Lu, L., Lu, S. Do social media trump government report cards in influencing consumer choice? Evidence from U.S. Nursing Homes. (2020). Available at SSRN: https://ssrn.com/abstract=3531964.
Luca, M., Smith, J., 2015. Strategic disclosure: the case of business school rankings. J. Econ. Behav. Organ. 112, 17–25.
Nosko C., Tadelis S. 2015. The limits of reputation in platform markets: an empirical analysis and field experiment. NBER Working Paper No. 20830, NBER
Perraillon, M.C., Konetzka, R.T., He, D., Werner, R.M., 2019. Consumer response to composite ratings of nursing home quality. AJHE 5 (2), 165–190.

Pope, D.G., 2009. Reacting to rankings: evidence from "America's Best Hospitals. J. Health Econ 28 (6), 1154–1165.
Reid, R.O., Deb, P., Howell, B.L., Shrank, W.H., 2013. Association between Medicare Advantage plan star ratings and enrollment. JAMA 309 (3), 267–274.
Werner, R.M., Konetzka, R.T., Polsky, D., 2016. Changes in consumer demand following public reporting of summary quality ratings: an evaluation in nursing homes. Health Serv. Res. 51, 1291–1309.
Tadelis, S., 2016. Reputation and feedback systems in online platform markets. Ann. Rev. Econ. 8, 321–340.