Navigating the Notches: Charity Responses to Ratings

Most Recent Version Here

Jennifer Mayo*

University of Michigan

November 4, 2021

Abstract

This paper studies both donor and nonprofit responses to the star rating system designed by Charity Navigator. Using IRS Form 990 data from 2002 to 2019, I find that an increase in a charity's rating from 3- to the highest 4-star rating is associated with a 6% rise in contributions, with larger effects among smaller charities. Some charities respond to the incentives by changing their behavior to try to get themselves above the star thresholds, leading to "bunching" at the thresholds. This response is equal to the effect of charities halving spending on administration. I find that some of the response is due to misreporting of expenses in order to achieve a higher star rating. The analysis suggests that a notched rating system induces greater behavioral change than a continuous measure, but affects a smaller number of charities. Which rating system is preferred depends on the relative value placed on these effects.

*Email: jenmayo@umich.edu. I am grateful for advice from Ashley Craig, Jim Hines, Nirupama Rao and Joel Slemrod, and for comments from Charlie Brown, Adam Cole, Vera Eichenauer, Edward Fox, Teresa Harrison, Stephanie Karol, Laura Kawano, Paul Kindsgrab, Shawn Martin, Jonathan Meer, Dylan Moore, Michael Murto, Andrew Simon, as well as seminar participants at the University of Michigan, WZB Berlin, the Online Public Finance Seminar, the 76th Annual Congress of the IIPF, and the NTA Annual Conference 2020.

1 Introduction

The U.S. nonprofit sector is a large sector in the economy, employing over 10% of the labor force, and attracting roughly $450 billion in donations every year. However, it is often difficult to monitor. With no profit maximization, and fewer accountability measures, the concern is that charities face weaker incentives to innovate and improve. In recent years, ratings websites and information aggregators have tried to change this by providing information to help consumers make better-informed decisions. However, any incentive structure gives rise to both intended and unintended consequences. In this paper, I focus on the star ratings system used by Charity Navigator and study both donor and nonprofit responses, attempting to disentangle real changes in behavior from manipulation.

Founded in 2001, Charity Navigator has become the most-utilized evaluator of charities in the Unites States, rating charities according to their financial health, accountability and transparency. In 2019, the site was visited by 8.5 million unique users over 11 million times[1]. Given that only the largest charities are rated, the number of users implies that Charity Navigator helps to direct up to $5 billion[2] worth of donations every year. Therefore, understanding the incentives created by the ratings system is of particular importance.

I find that the ratings system has its intended effect in redirecting donations to higher-rated charities: a one-star increase in rating from 3- to the highest 4-star rating raises total contributions the following period by 6%. The effect is even greater for smaller charities, which experience a 11% rise in contributions.

Charities respond to these incentives by taking steps to just exceed the star thresholds (i.e. "bunching"), especially the 4-star threshold. In particular, charities close to the thresholds score 2 percentage points higher than they would if they were to

[1] With approximately 55% of Americans giving to charity annually, this represents roughly 5% of givers.

[2] Average contributions of the 9,000 rated charities are $11 million, which means that almost $5 billion are at stake if 5% of American givers use Charity Navigator. This is an underestimate if donors discover charities' ratings through other means, e.g. the media or advertising campaigns.

be rated on a continuous measure. If organizations were to respond to ratings only by cutting administrative expenses, this would be equivalent to halving spending on administration. In practice, charities respond by reducing expenditure on both fundraising and administration. However, this response is not entirely real, with some charities misreporting or relabelling expenses in order to achieve a higher rating. Lastly, I argue that a notched rating system induces greater behavioral change from some charities than a continuous measure, but affects a smaller number of organizations. Which measure is preferred depends on the relative value placed on these effects.

The empirical strategy of this paper combines the bunching work pioneered by Saez (2010) with the "traces of evasion" literature (Slemrod and Weber, 2012). Previous work has documented the behavioral responses of firms to tax "notches" and "kinks" (Kleven and Waseem, 2013; Almunia and Lopez-Rodriguez, 2018; Velayudhan, 2019, etc.), but charities' responses have received much less attention. Recent notable exceptions include St. Clair (2016), Marx (2018) and Homonoff et al. (2020). In particular, Marx (2018) uses bunching methods to show that charities reduce revenues by up to $1000 in order to avoid reporting requirements, while Homonoff et al. (2020) suggest that 0.2% of public charities inflate their net assets by over $1,200 to appear solvent.

The analysis disentangles real changes in behavior from manipulation, both by using standard bunching estimation techniques to calculate the excess mass of nonprofits just above the star threshold, as well as by analyzing charities' financial information contained in the IRS Form 990. Exogenous changes to the rating methodology are used to estimate the donation responses, which can be considered as a charity's willingness-to-pay for an extra star. In particular, I exploit the introduction of "CN 2.1" in June 2016, which instigated three-year averaging of non-capacity metrics[3].

[3]Non-capacity metrics comprise Charity Navigator's financial efficiency metrics: program expense percentage, administrative expense percentage, fundraising expense percentage, and fundraising efficiency.

This paper is the first to examine the effect of ratings systems on the behavior of nonprofits. However, with information aggregators becoming increasingly popular, there is now a growing literature on the welfare impacts of websites such as Yelp and TripAdvisor (Lewis and Zarvas, 2016; Chen, 2018; Fang, 2019; Reimers and Waldfogel, 2020), as well as health sector ratings (Darden and McCarthy, 2015). Dranove and Jin (2010) provide a review of the literature on the theoretical issues related to quality disclosure.

Ratings are displayed prominently on the Charity Navigator website, and a higher star rating can affect donor and stakeholder decisions, and even managerial compensation. Elizabeth Ashbourne, the executive director of the Partnership for Quality Medical Donations, has called the rating "absolutely critical to attract new resources and donors," and many charities actively publicize their rating. As such, charities have an incentive to make decisions and adjust their financials to achieve a higher star rating[4].

By providing ratings, Charity Navigator helps to alleviate two potential market failures. The first relates to information, and the second to moral hazard. The information problem is such that, in the absence of Charity Navigator, donors are unable to observe charity quality and allocate resources according to their tastes and other easily observable charity characteristics[5]. This creates a mismatch between donors and charities, and without Charity Navigator, organizations are unlikely to provide this information - the "unravelling"[6] result, which predicts that organizations should reveal their quality voluntarily, is rarely observed in practice (Jin, 2005; Leuz, Triantis and Wang, 2008). Thus, by providing ratings and information on charities' financial health and governance and accountability practices, Charity Navigator helps donors to learn about charity quality, thereby improving efficiency.

[4] Accounting firm Wegner CPAs published a blog post on how to improve your Charity Navigator rating: https://www.wegnercpas.com/improving-rating-charity-navigator/.

[5] Krasteva and Yildirim (2013) establish, theoretically, that lowering the cost of information also has implications for fundraising strategy and the use of direct versus matching grants.

[6] "Unravelling" happens when the best quality firm discloses first as a way to differentiate themselves from lower quality competitors. Once the best firm discloses, the second-best firm faces the same incentive, and so on, until the worst firm discloses.

The moral hazard problem exists because charities' non-distribution constraint[7] prevents organizations from distributing operating profits to employees. This means that charity managers are unable to reap the rewards of cost-saving investments in the form of higher profits, thereby reducing pressure on managers to minimize costs. As a result, costs can rise unchecked at the expense of charitable services. Indeed, there are plenty of stories in the media of excessive spending[8], as well as lists of the worst offenders[9]. Charity Navigator's mission is therefore partly to change incentives to encourage nonprofits to reduce waste in return for a higher star rating.

Prior work has shown that donors value the financial health of the organizations to which they contribute. In particular, the degree of leverage, and the amount of cash holdings are negatively correlated with donations (Calabrese, 2011; Calabrese and Grizzle, 2012), but revenue from an organization's private sales do not generally crowd out contributions (Okten and Weisbrod, 2000).

Donor responses to charity ratings have also been studied. Brown et al. (2017) study responses to ratings in a laboratory setting, Harris and Nealy (2016) examine the effect of the three largest charity rating organizations, and Harris and Nealy (2021) analyze the impact of GuideStar's seals of transparency. Yörük (2016) focuses on the effect of Charity Navigator's rating system on contributions. Using a regression discontinuity design, he finds that a one-star increase in ratings increases donations to smaller charities by 20%. However, if charities that are better at improving their star rating are also better at attracting donations, then a regression discontinuity design will produce estimates that are biased upwards.

If charities are manipulating their financials in order to achieve higher ratings which, in turn, leads to increased donations, then this may lead to a reallocation of donations. These "extra" donations elicited by higher ratings might otherwise have been

[7] This is unrelated to distributions made by private foundations, which are not rated by Charity Navigator.

[8] Employees at the Wounded Warrior Project, for example, were reported to fly business class to events, and spend $500 a night on hotel rooms: https://nonprofitquarterly.org/wounded-warrior-project-the-fundraising-factory-issue/.

[9] https://medium.com/bestcompany/charity-scandals-the-six-ugliest-of-2016-19d3f1149a.

invested with different charities. Therefore, although the provision of information is welfare-enhancing, it is likely that welfare declines in the event of manipulation, or if charities are forced to divert resources away from charitable activities to improve ratings.

The rest of the paper proceeds as follows. Section 2 describes the Charity Navigator ratings system, captured in Section 3 in a simple framework. Section 4 presents the empirical estimates of the donation response, and Section 5 examines the charity response, including how much of the reported response is due to real changes in behavior versus manipulation. Section 6 discusses the lessons learned on how to optimally rate charities, and Section 7 concludes.

2 Data

2.1 IRS Form 990

The vast majority of data used by Charity Navigator are drawn from the IRS Form 990[10]. All IRS Section 501(c)(3) organizations[11] with gross receipts of over $50,000 are required to file this annual information return[12]. Failure to file for three consecutive years results in the automatic revocation of the organization's tax-exempt status.

The Form contains information on the charity's financials, as well as personnel and program service activities. In particular, Charity Navigator uses the following variables in its calculations: program expenses, administrative expenses, fundraising expenses, total expenses, total contributions, total assets and total liabilities. Program expenses comprise expenditures that relate to the charity's mission. Total

[10] As described in Section 2.2, charities must file a Form 990 in order to be rated.

[11] 501(c)(3) nonprofits are those whose mission relates to charity, education, science or public safety testing.

[12] Private foundations file a 990-PF, and churches and other houses of worship, as well as governmental organizations such as public universities are not required to file.

contributions include both private donations from individuals, estates, corporations, and/or other nonprofit organizations, as well as government grants (excluding reimbursements for services provided by the nonprofit under a government contract[13]). Fundraising includes both expenditures associated with fundraising events and professional fundraising fees[14].

Data on rated charities has been extracted from the Charity Navigator website (as described in Section 2.2), and information on unrated charities is provided by the Nonprofit Initiative for Open Data, which publishes the universe of IRS Form 990 e-filer data[15]. With a few exceptions[16], e-filing (as opposed to paper filing) is available to all organizations that file a Form 990, and required for those with net assets greater than $10 million. While the majority of the analysis in the paper focuses on rated charities, unrated charities are considered in Section 5.3.2.

2.2 Charity Navigator ratings

With over 1.5 million charities registered in the U.S., it is only possible for Charity Navigator to rate a subset[17]. The following guidelines are used to determine which charities to rate: the organization must (i) be registered as a 501(c)(3) public charity and file a Form 990; (ii) have generated at least $1 million in revenue for two consecutive years; (iii) have been in existence, with corresponding Forms 990 filed, for at least seven years; (iv) be based in the U.S., although its scope can be international; (v) have at least $500,000 in public support, and public support must account for at least 40% of total revenue for at least two consecutive years; (vi) have at least 1%

[13]These types of payments are reported as program service revenue.

[14]These are payments made to external organizations for conducting fundraising or for consulting on fundraising.

[15]These data are described here: https://github.com/lecy/Open-Data-for-Nonprofit-Research/blob/master/Build_IRS990_E-Filer_Datasets/Data_Dictionary.md.

[16]These include name change returns, short-year returns, returns from organizations with an exempt status application still pending, and returns older than the two prior years.

[17]Currently just under 9,000 charities are rated in a typical year.

of its expenses allocated to fundraising for three consecutive years; and (vii) have at least 1% of its expenses allocated to administrative expense for three consecutive years.

Rated charities are awarded a star rating based on an underlying score. These ratings are published roughly once a year, corresponding with the annual filing of the Form 990. Given that it takes time to calculate the rating[18], there is approximately a one-year lag between the release of the Form 990 data, and the publication of the rating. Accordingly, a rating published in 2019 usually corresponds to fiscal year 2018. Throughout the paper, the time period is defined as the publication year.

Ratings are clearly visible on the Charity Navigator website, appearing in the search results pages (see Figure 1). Highly-rated charities are also often identified by third parties as worthy causes[19]. Although the star rating is the most salient feature on the website, both donors and charities are able to find the underlying score, along with detailed descriptions of how the rating is calculated[20]. Charities are also able to initiate a rating review, or an appeal process, if they have any queries, or disagree with Charity Navigator's assessment[21].

Charity Navigator publicizes its ratings, and underlying metrics, via its Application Programming Interface (API)[22]. By scraping these data, I am able to collect information on the universe of rated charities, from 2002 to 2019. Using Charity Navigator's API, instead of extracting information from the IRS Form 990, also ensures that I use the identical line items used by Charity Navigator to generate the ratings.

The rating system employed by Charity Navigator has evolved over time. Initially, Charity Navigator scored organizations solely on the financial information provided

[18]Typically, Charity Navigator publishes the rating 3 to 4 months after a charity files its Form 990.

[19]For example, https://www.forbes.com/sites/jeffkart/2018/04/20/highly-rated-charities-to-consider-for-earth-day.

[20]See, for example, https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=1284.

[21]These are described in detail here: https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=6498.

[22]See https://charity.3scale.net/docs/data-api/reference.

in their annual Form 990 filings. In September 2011, a new dimension was added to the rating methodology called "Accountability & Transparency". This tried to capture measures of governance and was equally weighted with "Financial Health". Finally, on June 1st 2016, Charity Navigator launched its current methodology ("CN 2.1")[23], which changed the Financial Health metric by introducing three-year averaging of non-capacity metrics and adding a new component, the Liabilities to Assets Ratio.

The overall score depends equally on the "Financial Health" and "Accountability & Transparency" scores, and is determined by the following formula:

$$100 - \sqrt{\frac{(100 - Financial)^2 + (100 - A\&T)^2}{2}} \tag{1}$$

Such a formulation means that charities are more likely to earn a high star rating if they excel in both areas. Raw overall scores are out of 100, but stars are awarded according to designated thresholds, as follows[24]:

Overall Rating:	★★★★	★★★☆	★★☆☆	★☆☆☆	0 Stars	Donor Advisory
Overall Score:	≥ 90	80 - 90	70 - 80	55 - 70	< 55	N/A

From above, the overall score is clearly dependent on the two scores. The "Financial Health" score is derived from information provided on the IRS Form 990, while the "Accountability & Transparency" score uses both Form 990 data and also information available on the charity's website. Both scores rely on publicly available information, but the value comes from synthesizing available financial data in order to provide donors with easily accessed and digestible information.

[23] In August 2020, Charity Navigator announced their intention to change the methodology again, and expand rating to over 150,000 charities a year.

[24] Although the thresholds occur at round numbers, the overall score is a complex function of several metrics, meaning that round number bunching is not a concern.

2.2.1 Financial Health

The "Financial Health" score is comprised of seven financial metrics: (i) Program Expense Percentage; (ii) Administrative Expense Percentage; (iii) Fundraising Expense Percentage[25]; (iv) Fundraising Efficiency[26]; (v) Program Expenses Growth; (vi) Working Capital Ratio[27]; (vii) Liabilities to Assets Ratio. A higher value of (i), (v) and (vi) earn a charity a higher score, while a higher value of (ii), (iii), (iv) and (vii) cause a lower score. Each metric is graded out of 10, with the overall "Financial Health" score calculated by combining them and adding 30 points (to convert the scores to a 100 point scale). Other than (i) and (v), which are continuous measures, a charity can earn either 0, 2.5, 5, 7.5 or 10 points depending on the value of a metric. It is important to note that, for each financial metric, the conversion that is applied to move from the raw score to a 10-point scale varies by sector[28]. For example, community foundations, food banks, food pantries & food distribution, and humanitarian relief supplies must achieve a program expenses ratio of at least 92% to be awarded 10 points, whereas museums need achieve only 83% for a perfect score. Charity Navigator explains that this is because "different types of charities have different resource and spending requirements".

Finally, in contrast to previous versions of the methodology, which simply assessed data from the most recent fiscal year, under "CN 2.1", the first four financial metrics are now calculated by averaging data from the charity's three most recent fiscal years. Program Expenses Growth is calculated using data from the charity's three to five most recent years[29], while metrics (vi) and (vii) use just the most recent information.

[25] Each of these percentages use total expenses as the denominator.

[26] This is calculated as $\frac{\text{fundraising expenses}}{\text{total contributions}}$.

[27] This is calculated as $\frac{\text{unrestricted net assets}}{\text{total expenses}}$.

[28] For a full description of financial score conversions, see the Charity Navigator website https://www.charitynavigator.org/.

[29] In most cases, Charity Navigator computes annualized growth over the four most recent fiscal years. However, if an organization has engaged in non-recurring, atypical activities in the first of the four years, then five years are evaluated. If a fifth year is unavailable, the window is reduced to three years.

2.2.2 Accountability & Transparency

The "Accountability & Transparency" score aims to capture governance practices and information accessibility. Contrary to the calculation of the "Financial Health" score, each charity starts with a base score of 100, and deductions are made for failing to meet 17 criteria[30], 12 of which are derived from the IRS Form 990. For example, points are deducted if charities do not have an independent board or conflict of interest policy. The remaining five criteria relate to the charity's website. For example, points are deducted if board members are not listed on the website or if the most recent Form 990 is not published.

2.3 Summary statistics

Charity Navigator currently rates roughly 9,000 charities a year. Most charities earn 3 or 4 stars, with the proportion earning the highest rating increasing over time[31]. Table 1 presents some summary statistics for the rated charities, and Appendix Table B.2 shows the distribution of rated charities by sector. Arts, Culture and Humanities, and Human Services charities are the most rated, representing 15% and 13% of the sample, respectively. Relative to all 501(c)(3) charities, Charity Navigator rates more Arts, Culture and Humanities charities, and far fewer Education charities[32]. Given the ratings criteria described in Section 2.1, it is unsurprising that Table 1 shows that rated charities are relatively large, with revenue of almost $15 million, on average. Expenses are close to $14 million, with program expenses accounting for roughly 85% of the total. Administrative and fundraising expenses make up the rest, accounting for 8% and 7%, respectively.

[30] The amounts deducted are the same regardless of sector - for a full list, see Appendix Table B.1.

[31] Between 2002 and 2019, 33% of charities earned 4 stars, and 43% earned 3 stars. These numbers increase to 39% and 46%, respectively, if we just consider the most recent period, 2017-19.

[32] 18% of all 501(c)(3) charities are Education charities, but only 6% of rated charities operate in the Education sector.

3 Conceptual framework

The following section provides a framework for understanding the need for Charity Navigator, and how the ratings system affects an organization's choices[33].

Charity Navigator helps to minimize a potential matching problem for donors and a moral hazard problem for charities. Donors derive utility from charity quality and idiosyncratic tastes. However, they are unable to observe charity quality and, without Charity Navigator, would allocate resources according to their tastes and other easily observable charity characteristics, such as sector and location. This lack of information creates a mismatch between efficiency and resource allocation. By providing ratings and information on charities' financial health and governance and accountability practices, Charity Navigator helps donors to learn about charity quality, thereby improving efficiency[34].

On the charity side, the issue is one of moral hazard. Unlike for-profit firms, which reap the rewards of cost-saving investments in the form of higher profits, charity managers face a non-distribution constraint, meaning that profits must be reinvested or can be enjoyed as "perquisites" (lower effort levels, shorter work days, more generous benefits, etc). Given that these in-kind perquisites are less valuable than cash, charity managers have less of an incentive to exert effort on implementing cost-saving measures[35]. As a result, costs can rise unchecked at the expense of charitable services. By introducing a ratings system, Charity Navigator strengthens the incentive to cut costs, and eases the moral hazard problem[36].

The framework outlined below focuses on the way the ratings system affects a charity's choice of inputs. Under the ratings system, charities seek to maximize charitable services, but face a trade-off between improving their score and investing in

[33] I take the ratings system itself as given, and do not offer any judgement on the criteria.

[34] Of course, a first-best solution would be if we knew exactly how much "good" each charity is doing. Charity Navigator's rating system offers a second-best solution.

[35] See Glaeser and Shleifer (2001) for a discussion of for- versus non-profit entrepreneurs.

[36] In the longer run, the rating system could alter a charity's mix of donations versus earned income, and even affect their choice of service provision. However, I save this for future work.

inputs that reduce their score but are complementary to the production of these services.

Organizations choose their inputs, a, which include metrics such as the ratio of administrative to total expenses. Charities also have the option to misreport (or avoid), where reported inputs are denoted by $\hat{a}$, and the amount of misreporting is represented by $m := (a - \hat{a}) \geq 0$. Charities are heterogeneous in a productivity parameter, ψ, which is exogenously distributed over the range $[\underline{\psi}, \overline{\psi}]$.

The cost of producing charitable services is denoted by $C(a, m, \psi)$, and includes the cost of inputs, $B(a, \psi)$ (the cost of organizing a fundraising event, for example), as well as the cost of avoidance, $A(a, m, \psi)$[37]. $C(a, m, \psi)$ is nondecreasing and convex in a and m, and decreasing and convex in ψ. Note that there are no externalities in the model, which means that the marginal benefits of real and avoidance costs are equated[38]. Appendix A demonstrates this using a quadratic cost function.

Lastly, charitable services, CS, are produced according to the production function $g(a, D, O)$, where donations, D, are increasing in a charity's score (s), which is itself negatively correlated with (reported) inputs: $D := D(s(\hat{a}))$. O stands for other sources of revenue, such as business income and sale of inventory. $g(a, D, O)$ is assumed to be strictly continuous, increasing and concave in each argument. Accordingly, inputs, such as administration and fundraising, aid in the production of charitable services but also reduce donations.

Charities choose a and m to maximize the production of charitable services:

$$CS := g(a, D, O) - C(a, m, \psi)$$

We can consider an organization's choice of a and m under three different scenarios: (i) in a world without Charity Navigator; (ii) when the rating is a continuous measure

[37] ψ is unrelated, but possibly correlated with, the cost of avoidance, which includes both financial and psychic costs.

[38] The cost of misreporting is the same as the cost of paying an accountant to help you really reduce your inputs.

(the underlying score); (iii) when the rating is discrete (the star rating).

If the ratings system did not exist, charities would have no score concerns, and donations would just be a function of $\hat{a}$, and not s. In this case, optimal avoidance would be zero ($a = \hat{a}$)[39], and donations would not be negatively affected by expenditure on inputs. Indeed, higher expenditure on administration and fundraising may increase donations[40]. The choice of a would therefore be characterized by the following first-order condition:

$$g_a(a, D, O) + g_D(a, D, O)D_a = C_a(a, m, \psi)$$

where the left-hand side of the equation is the marginal benefit of an increase in input a, and the right-hand side is the marginal cost. This first-order condition defines charities' preferred level of input expenditure, $a^*(\psi)$. If $\frac{da^*}{d\psi} = -\frac{B_{a\psi}(a,\psi)}{B_{aa}(a,\psi)} > 0$, then there will be a one-to-one relationship between a^* and ψ, implying the inverse function $\psi(a^*)$. Henceforth, I define $C(a, m, a^*) = A(a, m, \psi(a^*)) + B(a, \psi(a^*))$.

However, if charities are rated using a continuous measure, then donations are again a function of the underlying score, and charities choose a and m to maximize the production of charitable services such that the first order conditions are as follows[41]:

$$g_a(a, D, O) = g_D(a, D, O)(D_s \cdot s_a) + C_a(a, m, \psi)$$

$$g(a, D, O)(D_s \cdot s_m) = C_u(a, m, \psi)$$

The marginal benefit of increasing expenditure on a comes from the increase in the production of charitable services. This is weighed against the increased cost of

[39] I abstract from the possibility that charities are motivated to misreport for the benefit of donors or other stakeholders in the absence of the rating system.

[40] If donors care about expenditure on non-program services, then $D_a > 0$ but $D_{aa} < 0$.

[41] Similar distortions would arise if the model featured reputation concerns, diversion of resources to principles, or performance-pay contracts.

inputs, as well as the negative effect on donations (via the score). Note again that the welfare costs associated with increasing a relate to both selection ($g_D(a, D, O)(D_s \cdot s_a)$) and moral hazard ($C_a(a, m, \psi)$). If charities misreport, they weigh the increase in donations against the marginal cost of avoidance, $C_u(a, m, \psi)$.

The third scenario is one where charities are rated using a discrete measure. This reflects Charity Navigator's rating system, in that donors do not observe the underlying score. Instead, they observe a star rating, s, which is a discontinuous function of the underlying score. In other words, the star rating introduces a notch into the system, so that when $\hat{a}$ falls below $\overline{a}$, the star rating jumps from s_0 to s_1, where $s_1 > s_0$[42]:

$$s = s_0 \quad \text{if} \quad \hat{a} > \overline{a}$$

$$s = s_1 \quad \text{if} \quad \hat{a} \leq \overline{a}$$

If charities score s_1, they receive a boost in donations, denoted by $\delta(\psi)$, where $\delta_\psi > 0$.

We can therefore re-write the charity's problem as:

$$\max_{a,m} \{g(a, D, O) - C(a, m, \psi)\}$$

but now $D = D_0$ if $a^* > \overline{a}$ and $D = D_1 = D_0 + \delta(\psi)$ if $a^* \leq \overline{a}$.

When will a charity change its behavior to exceed the threshold, $\overline{a}$? If $a^* \leq \overline{a}$, there is no need to misreport and a charity will report expenses truthfully as a^*. Re-writing $\delta(\psi)$ as $\delta(a^*)$, if $a^* > \overline{a}$ then the charity earns $g(a^*, D_0, O) - C(a^*, 0, a^*)$ if it reports $\hat{a} > \overline{a}$, and $g(a, D_1, O) - C(a, a - \overline{a}, a^*)$ if it reports $\hat{a} = \overline{a}$[43]. Note that the degree of

[42] In reality, Charity Navigator's rating system features four star thresholds. For clarity, I just assume one, but the predictions would be unchanged if charities faced four thresholds, as long as charities do not jump over more than one threshold at a time.

[43] Production is denoted $g(a, D_1, O)$ when $\hat{a} = \overline{a}$ to allow for both real and avoidance responses. If the response was purely avoidance, charities would continue to use inputs a^*.

bunching is affected by the number of people who use Charity Navigator - the more users there are, the greater the effect of the score on donations. Given that costs are convex, we can define $\gamma(\overline{a}, \delta, a^*)$ as the maximum difference between a^* and $\overline{a}$ that a charity would be willing to choose $\hat{a} = \overline{a}$. In other words, a charity bunches if $\overline{a} < a^* \leq \overline{a} + \gamma(\overline{a}, \delta, a^*)$. Thus, reported expenditure on inputs can be characterized as follows:

$$\hat{a} = \begin{cases} a^* & \text{if } a^* \leq \overline{a} \\ \overline{a} & \text{if } \overline{a} < a^* \leq \overline{a} + \gamma \\ a^* & \text{if } a^* > \overline{a} + \gamma \end{cases}$$

The framework therefore highlights that the amount by which charities reduce reported inputs depends on the value of a higher star, and that the degree of misreporting depends on the cost of avoidance (this is discussed in Section 5.3.1). Lastly, under a continuous measure, all charities face an incentive to reduce their reported inputs. However, under a discrete measure, only those close to the threshold will be induced to change their behavior.

4 Donation responses

The model presented in Section 3 highlights that any behavioral responses by the charity will be a reflection of the trade-off between improving an organization's score and furthering its mission[44]. Therefore, in order to fully understand the charity response, we also need to understand the donor response. The observed charity response is dependent on the value of an extra star, because charities will only respond to the rating system if they are rewarded for doing so. The donor response is the reward for a higher star that I am able to measure. However, even if the donation response is zero, we might still expect a charity response, as a higher star

[44]Increased spending on program services is the only thing that both improves a charity's score and expands production of charitable services.

rating is likely to be valued by other stakeholders, such as the board of directors or grant-making organizations. However, if a higher star rating elicits much higher donations, then a charity's willingness-to-pay for an extra star increases.

The first step is to check that it is indeed the star rating that is important, and not the underlying score. If donors care about the score, then there is less of an incentive for charities to bunch at the star thresholds. Appendix Table B.3 presents results of a fixed effects regression showing that a 1-point improvement in the underlying score (without crossing a threshold) is associated with a 0.8% increase in contributions. This compares to almost 8% when the star rating increases, which shows that crossing a star threshold is a lot more valuable than improving the underlying score. However, scores and stars are not randomly assigned - there are likely to be unobservable characteristics, such as managerial quality, which affect both donations and the star rating. In particular, if good managers are associated with both increased donations and a better rating, then these estimates will be biased upwards. In order to move closer to a causal interpretation, ratings endogeneity must be accounted for.

Using a regression discontinuity (RD) framework, Yörük (2016) finds that a one-star increase in ratings results in a 14.5% rise in contributions the following period. However, given the possible manipulation of scores, a standard RD approach is likely to be inappropriate[45]. In particular, if charities with greater capacity to adjust their financials to bunch at the thresholds are also better at attracting donations, then estimates will be upwardly biased. Indeed, a formal test of manipulation (Cattaneo et al., 2018) rejects the null hypothesis of no discontinuity in the density of observations at the star thresholds. Therefore, in order to accurately measure the donor response, we need an exogenous change in a charity's star rating.

As mentioned in Section 2, "CN 2.1" launched on June 1st 2016, introducing three-year averaging of non-capacity metrics, as well as a new capacity metric, Liabilities to Assets Ratio. As a result of this change, around 27% of charities received a new star rating overnight: 19% received a 1-star increase and 8% a 1-star decline. By

[45]See, for example, Lee and Lemieux (2010) for a discussion.

studying the charities that experienced an exogenous change in their star rating, we can derive an unbiased estimate of the effect on donations. The key assumption here is that the change in methodology was a surprise[46], and that charities were unable to adjust their metrics in anticipation. Given that CN 2.1 was based on averaging historic metrics, adjustments were impossible.

Using the period 2015-2018, I estimate a similar model as Yörük (2016), but exploit the exogenous change in star rating:

$$ln(Contributions)_{it} = \alpha RatingIncrease_{it-1} + \beta' X_{it-1} + \lambda_i + \tau_t + \epsilon_{it} \qquad (2)$$

where the dependent variable is the natural logarithm of total contributions received by charity i in time t in 2010 dollars. The main explanatory variable, $RatingIncrease_{it-1}$, is a dummy variable equal to 1 if the charity received a higher star rating under the new methodology. X_{it-1} is a vector of controls, including fundraising expenses, total net assets and fundraising efficiency ($\frac{\text{fundraising expenses}}{\text{total contributions}}$). λ_i and τ_t are charity and time fixed effects respectively, and ϵ_{it} is an error term. Without exploiting the methodological change, the concern is that $\mathbb{E}(\epsilon_{it} \mid RatingIncrease_{it-1}) \neq 0$. In other words, unobserved factors that affect a charity's ability to solicit donations may be systematically correlated with its ability to increase its rating.

I run two separate versions of the model. The first compares charities whose 3-star rating remained unchanged under CN 2.1 with charities whose rating increased from 3 to 4 stars. The second version compares 2-star charities with those whose rating increased to 3 stars under CN 2.1. Table 2 presents the results for the former, and Table 3 the latter. Given that the data are collected at the charity level, it is impossible to know how donor characteristics affect the response. It might be the case that sophisticated donors realize that the star increase is just the product of a methodological change, and not a real change in charity behavior or strategy.

[46] Newspaper articles from the time seem to suggest that the change was indeed a surprise. See, for example, https://www.bkd.com/article/2016/07/charity-navigator-releases-new-rating-methodology.

Therefore, the estimated coefficients represent the donation response averaged over sophisticated and unsophisticated donors[47].

The estimated coefficients in Table 2 imply that a one-star increase in rating from 3 to 4 stars raises total contributions the following period by 6%. The effect is even larger for smaller charities, which experience a 11% rise in contributions. Large charities see a 9% rise in donations, and medium-sized charities gain the least[48]. This non-monotonicity in effect sizes is likely due to the fact that a rating for a small charity contains more information than a rating for a charity that is already well-known to donors. The response for large charities may be partly driven by publicity.

Similarly, for charities that experience an exogenous increase in rating from 2 to 3 stars, the coefficients presented in Table 3 show that these organizations see donations rise by 8%. Small and medium-sized charities experience an 8% rise, but large charities see no effect. These effect sizes are significantly larger than the impact of media campaigns, which appear to have no effect on giving (Yörük, 2012). However, they are very similar to the effect of a one-star increase in Yelp ratings on restaurant revenues (Luca, 2011)[49].

These results can also be compared to Yörük (2016), who recall finds that a one-star increase from 3 to 4 stars is associated with a 14.5% rise in contributions and, pooling all star increases, estimates that small charities experience a 19.5% increase in contributions. These coefficients are larger than those presented in Tables 2 and 3[50]. This is to be expected if the charities manipulating ratings are also the ones that are "better" at soliciting donations. In other words, without exploiting the exogenous change in ratings induced by the introduction of CN 2.1, we might expect that $\mathbb{E}(\epsilon_{it} \mid RatingIncrease_{it-1}) > 0$, meaning that coefficients estimated via RD are

[47]For this reason, and the fact that the rating is almost certainly valued by other stakeholders, these estimates likely understate the reward for a higher star.

[48]A small charity is defined as having net assets less that $1 million and a large charity as greater than $57 million.

[49]Implementing an RD design around the rounding thresholds, Luca (2011) finds that a one-star increase in Yelp rating leads to a 5-9 percent increase in revenue.

[50]Yörük (2016) studies the time period 2007-2010, when Charity Navigator was less popular, and bunching on the charity side was less sharp.

biased upwards.

The results presented in Tables 2 and 3 suggest that, if all charities are considered, an increase in star rating from 2 to 3 stars is more valuable than an increase from 3 to 4 stars. This may be because only 4 stars are available, so increasing a rating from 2 to 3 stars moves you from the "bottom" half of charities to the "top". However, if we focus on smaller charities, an increase from 3 to 4 stars is more valuable. This is unsurprising if Charity Navigator's role is to provide information to potential donors. Information on smaller charities is more revealing, as their financials are perhaps under less scrutiny and they are less exposed to media coverage.

4.1 Symmetry in response

If an increase in star rating generates a rise in donations, does a fall produce a reduction in donations? To answer this question, I next repeat the exercises of Section 4 but compare charities whose 4-star rating remained unchanged under CN 2.1 with charities whose rating decreased from 4 to 3 stars, and 3-star charities with those whose rating decreased to 2 stars under CN 2.1. Tables 4 and 5 present the results, which show that all the estimated coefficients are statistically insignificant from zero[51]. In other words, I can not reject that earning a lower star does not significantly reduce donations[52]. One explanation for this might be that donors are slow to update their beliefs downwards[53]. Once they see that a charity is of a certain quality, they do not continue to check the rating on Charity Navigator. This could suggest that the asymmetry is driven by new donors: existing donors continue to support the charity, regardless of the downrating, while new donors choose to donate to 4-star charities instead. However, without donor-level information, it is

[51]The coefficient in Table 4 is marginally insignificant for large charities, suggesting that some donors punish these organizations for losing a 4-star rating.

[52]The only sector where this does not seem to be the case is among Performing Arts organizations. This is consistent with Mayo (2021), who finds that donors view some of these organizations as substitutes for each other.

[53]It's also possible that charities convinced donors that any downgrade was just a product of a formulaic change.

not possible to test this hypothesis.

This raises the question as to whether the ratings system creates a zero-sum game. In other words, is one charity's gain in donations (induced by Charity Navigator) another charity's loss[54]? This is not the perfect setting to study whether donations are in fixed supply, but examining how the methodological change affects total donations to all charities offers some suggestive evidence. Just focusing on donations in the period immediately prior to the methodological change, and immediately post, total donations increased from $99,782 million to $105,780 million (in 2010 dollars). Restricting attention to 3- and 4-star charities, these figures are $92,784 million and $97,661 million respectively. Restricting the sample further to charities whose rating either remained at 3 stars or increased to 4 stars, the increase in donations is from $33,280 million to $35,447 million. Therefore, these data seem to suggest that the rating system did not create a zero-sum game among 3- and 4-star charities.

However, we can also examine the impact of the methodological change on 1- and 2-star charities, as well as unrated organizations. Contributions to 1-star charities were unchanged, and the share of total donations allocated to rated versus unrated charities also remained the same. However, total donations to 2-star charities fell from $17,021 million to $13,337 million. This suggests that some of the extra donations being awarded to newly-rated 3- and 4-star charities came at the expense of 2-star charities. Again, this is consistent with donors viewing charities as in either the "top" or "bottom" half of the distribution.

4.2 Robustness

Although the change in the methodology came as a surprise to charities, two concerns remain regarding the OLS estimation: first, that charities that received a new star rating are a selected sample, and second, that by pooling the post-period, charities have time to react to the new ratings system such that the estimated coefficients

[54]Deryugina and Marx (2021) finds that donations in response to deadly tornadoes do not come at the expense of other charities, suggesting that giving need not be in fixed supply.

represent both the donor response, and a behavioral response from the charity. I address each in turn.

Charities that receive a new star rating are remarkably similar in observable characteristics to those whose ratings remained unchanged. In particular, I test whether the two groups are of a similar size and operate in similar sectors. I find that changes in star ratings are evenly distributed across sectors, and that there is no significant difference in total assets or contributions between the two groups prior to the ratings change. I also examine which charities would have received a new rating had the change in methodology taken place in 2015 instead of 2016. Again, I find that those that would have experienced a rating change in 2015 are not significantly different from those whose rating would have remained the same.

As a final check, I employ an IV estimation strategy that involves calculating charities' scores, assuming that 3-year averaging had been in place throughout the entire time period. I then use the charities that would have seen their star rating change in June 2016 if averaging had already been in place as instruments for those that actually saw their star rating change. The estimated coefficients presented in Appendix Table B.5 are very similar in magnitude to those estimated via OLS, but are insignificant. This is because, compared to the actual number of newly-rated charities, fewer charities "receive" a new rating under 3-year averaging[55], causing less variation in the key dependent variable. Despite the lack of power, the similarity in the coefficient suggests that the OLS estimates are indeed capturing the donor response, and not the charity's. These exercises are reassuring in that the organizations receiving a new rating in 2016 are not endogenously selected.

The other concern is that, by pooling the post-period (June 2016 to 2018), charities have time to respond to the methodological change, and that the coefficients estimated via OLS at least partially reflect this response. In other words, the OLS regressions may not isolate the donor response. The lag between a charity filing Form

[55]Charities that receive a higher star rating when I recalculate the pre-period using 3-year averages are charities that are hovering around the threshold.

990 and the ratings publication date makes this unlikely[56]. However, I conduct several robustness checks in order to allay this concern. The first is to omit 2018[57], and run the regressions using the time period 2015-2017. Given that the ratings change was introduced in June 2016, omitting 2018 gives charities only one year to respond to the change, and any response is weighted by a third (due to three-year averaging). From these robustness checks, I find the coefficients to be quite stable (see Appendix Table B.4), suggesting that the response is driven by donors, not charities.

As another check, I use a second IV estimation strategy, "fixing" a charity's rating at the rating it receives in the immediate post period, not allowing for any further changes. Again, I use these ratings as an instrument for the actual ratings. Here, the estimated coefficients presented in Appendix Table B.6 are larger than the coefficients estimated using OLS, but follow the same pattern. This difference occurs because the instrument identifies charities that are consistently "better" or "worse" (ignoring those whose rating fluctuates) once the methodology changes. In other words, the instrument isolates charities whose rating changed when the methodology changed, and then remained at the new rating they were awarded. Indeed, these results reveal that the donation response is stronger for charities that consistently earn 4 stars.

Finally, the distribution of charities' overall scores, depicted in Figure 3, shows that organizations do not appear to bunch at the 4-star threshold from 2016 to 2018. This implies that there is little charity response during this time period, and that the OLS estimates reflect a pure donor response.

In conclusion of this section, these exercises are reassuring that the charities receiving a new star rating are indeed representative of all rated organizations, and that the donation responses estimated via OLS are not contaminated by any behavioral

[56]On average, there is a one-year gap between a charity's fiscal year and the publication year.

[57]Ratings published in 2018 use Form 990s that reflect charities' 2017 fiscal year. A 2017 fiscal year can end any time from December 2017 to November 2018. Roughly 60% of charities' fiscal years coincide with the calendar year, and most have either June or December fiscal year ends. This delay in publication date also means that the Tax Cuts and Jobs Act, which passed in December 2017 and reduced individuals' incentive to give, does not affect the results.

response from the charity. In particular, the results show that a one-star increase in rating from 3- to 4-stars increases donations the following period by 6%, with even larger effects for smaller charities. The effect is asymmetric in that charities that see their rating decline do not experience a subsequent fall in donations. However, the observed increase in donations is at least partly at the expense of 2-star charities, which do experience a decline in contributions. Lastly, these results are almost certainly an underestimate of the gains to a higher rating, as they ignore any benefits such as improved reputation among grant-makers or the approval of board members. Charity managers and employees may also gain if salaries are tied to performance metrics.

5 Charity responses

Having established that charities are rewarded for a higher star rating, we can now examine the ways in which they respond. Section 5.1 presents descriptive evidence of a behavioral response, which Section 5.2 quantifies using bunching techniques. Lastly, Section 5.3 examines whether the observed response represents real changes in expenditures, or deliberate misreporting.

5.1 Descriptive analysis

5.1.1 Bunching in the overall score

If charities respond to the the ratings system, then the distribution of scores should feature "extra" observations on the preferred side of the notch. Excluding the years the rating methodology changed (2011 and 2016), Figure 2 illustrates the distribution of the overall scores relative to the star thresholds. Each notch point is represented by a vertical dashed line and is itself part of the preferred side of the notch. A charity with a raw score of 90 and above is awarded 4 stars, between 80 and 90 is 3 stars, 70-80 is 2 stars, 55-70 is 1 star, and less than 55 is no stars. The histogram displays

greater mass just to the right of the 2-, 3- and 4-star thresholds. For example, there are 6,697 observations[58] with scores between 89 and 90, but 7,644 with scores between 90 and 91. This suggests that some charities are exerting effort to cross the thresholds, especially to achieve 3 and 4 stars.

Given that Charity Navigator has become increasingly popular, we can also examine how the distribution of overall scores has evolved over time. Figure 3 again depicts the distribution of the overall scores, but focuses on the most recent 4 years (2016-2019). When Charity Navigator changed its methodology (on June 1st 2016) this resulted in a more diffuse bunching pattern that year. However, once charities became familiar with the new rating system, the bunching pattern became more pronounced, especially at the 4 star threshold. This is consistent both with charity learning and Charity Navigator's growing influence, being featured in media outlets such as Forbes and the New York Times.

Finally, we can examine heterogeneity by sector, size, use of a tax preparer, and reliance on contributions. The degree of bunching is found to vary by sector, with charities operating in Medicine and the Arts, Culture and Humanities sectors not bunching at all. Food Banks and those providing Human Services and Housing and Shelter bunch strongly at the 4-star threshold. There are several possible, related explanations. First, charities that fund disease research often have very loyal donors, who may pay less attention to ratings. Second, charities operating in the Arts, Culture and Humanities sector are less reliant on contributions for revenue and therefore face weaker incentives to gain a higher rating. Third, charities such as food banks and those providing human services or housing, often operate in very competitive markets - these charities provide fairly homogeneous services, so competition for donors is fierce, and earning a higher star rating is one way that organizations can differentiate themselves.

Classifying a small charity as one with net assets less that $800,000 and a large charity as having net assets greater than $47.5 million (corresponding to the 10th and 90th

[58] An observation is a charity, year pair.

percentile, respectively), Appendix Figure C.1 depicts the distribution of the overall score by charity size (2017-19). It shows that small charities bunch more at the 2- and 3-star thresholds, while larger charities bunch more at the 4-star threshold (but not at all at the 3-star threshold). This is unsurprising if larger charities have more resources to devote to understanding the ratings system, and improving their score. This pattern is also broadly consistent with donation responses, which were found to be stronger for smaller charities. Indeed, if small charities gain the most from increasing their star rating, then they should be the most "willing-to-pay" for an extra star.

Appendix Figure C.2 examines heterogeneity by tax preparer[59]. Because Charity Navigator rates only the largest charities, and organizations lose points if their accounts are not audited, it is unsurprising that over 85% of rated charities employ a tax preparer. Appendix Figure C.2 compares the bunching pattern of organizations employing tax preparers against those without tax preparers, and examines whether there is further heterogeneity by those that employ an experienced tax preparer (classified as those that have prepared accounts for rated charities more than 100 times[60]), or a preparer with a CPA license. Major differences across preparer status or types are not apparent. Experienced preparers and those with a CPA license appear to be helping organizations to bunch more at the 3-star versus 4-star threshold. This suggests that the bunching at the 4-star threshold is being driven by less experienced preparers, those without a CPA license, and charities with no preparer at all.

Lastly, bunching behavior can be examined by reliance on contributions, with the expectation that charities most reliant on contributions for revenue are more likely to value a higher star rating. This is confirmed in Appendix Figure C.3, which presents the distribution of the overall scores for charities most reliant (90th percentile) and least reliant (10th percentile) on contributions (2017-19)[61]. The analysis suggests that charities most reliant on contributions bunch more at the 4-star threshold. The

[59]Note that this information only exists starting in 2010 for charities that e-file.

[60]This corresponds roughly to the 90th percentile.

[61]The least reliant charities earn less than 42% of their revenue from contributions; the most reliant earn at least 99% from contributions.

least reliant charities, on the other hand, tend to bunch at the 3-star threshold, but not at all at the 4-star threshold.

5.1.2 Bunching in metrics

Given the observed distribution of the overall scores, it is worth asking which component metric is driving this pattern? Following the introduction of the Accountability and Transparency score in 2011, we can decompose the overall score into its component parts: Financial Health and Accountability and Transparency (A&T). A score of 90 in Financial Health or A&T does not automatically earn a charity 4 stars (the star rating depends on the overall score), but it can boost the overall score. Furthermore, charities may wish to advertise individual scores. For example, a charity that earns 3 stars overall, but 4 stars for Financial Health, may choose to publicize this.

Figure 4 presents the distribution of the Financial Health and A&T scores. Focusing on the 4-star threshold and the Financial Health score, Figure 4 shows charities to be bunching to the right of the highest Financial Health threshold. However, if the A&T scores are examined, the same pattern is not apparent. Instead, there is a mass of charities earning 89 points. Appendix Table B.1 shows that charities can boost their A&T score by meeting certain criteria, and can lose either 3, 4, 7 or 15 points depending on which criteria are unmet. Given the mass at 89, it must be that charities are not exerting as much effort to improve their A&T score. In other words, they are accepting the loss of 11 points, when they could be losing fewer if they were to implement some of the Charity Navigator-recommended policies.

The pattern of bunching in the Financial score prompts an examination of its component parts. As with the overall and Financial Health scores, most financial metrics are also notched. Depending on the value of a ratio (which varies by sector), a charity can earn either 0, 2.5, 5, 7.5 or 10 points. So again, charities have an incentive to improve each ratio by just enough to place them in the higher point bracket. There is little evidence of bunching in the capacity metrics, but Appendix Figure C.4 presents

the distributions of Administrative Expense Percentages for the sectors that appear to bunch the most on the overall score[62]. Here we see some evidence of bunching at the highest threshold, especially in the very competitive Food sector. The two fundraising metrics also display slight evidence of bunching, but the patterns are not quite as strong. The absence of bunching in financial metrics among unrated charities is notable.

5.2 Bunching estimation

5.2.1 Empirical strategy

Having identified a theoretical reason for bunching, as well as descriptive evidence of it, we now turn to estimation. In order to quantify the behavioral response to the rating system, I estimate the excess mass just above the star thresholds using bunching techniques first pioneered by Saez (2010). These methods were first used to examine individual and firm responses to taxation (Chetty et al., 2011; Kleven and Waseem, 2013; Almunia and Lopez-Rodriguez, 2018; Velayudhan, 2019), and more recently, to study income manipulation by nonprofits (St. Clair, 2016; Marx, 2018; Homonoff et al., 2020).

In my setting, estimation of the excess mass requires constructing a counterfactual distribution of scores in the absence of the star thresholds, and comparing this with the observed distribution. To do this, I collapse the data into counts of charities within score bins of 0.2 points and estimate the counterfactual density by fitting a 4th degree polynomial[63] to these counts, as follows:

$$F_k = \sum_{i=0}^{4} \beta^i S_k^i + \sum_{k=s^{lb}}^{s^{ub}} \omega_k \mathbb{1}(S_k = k) + \epsilon_k \tag{3}$$

[62] A lower ratio earns you more points, with the nominal standards highest for Food and Humanitarian Relief charities, and lowest for Museums.

[63] Results are insensitive to the choice of the order of polynomial, between a 3rd and 6th degree.

where F_k is the actual density of charities in each score bin, k, S_k is the midpoint of the score in each bin, and β^i is the coefficient on the polynomial terms. The ω_k identify either the missing or excess mass within each score bin relative to the counterfactual density, and ϵ_k is an error term.

The upper bounds, s^{ub}, are set visually, at the points where the density appears to show a discontinuous change[64]. I iterate over different choices of the lower bound s^{lb} to find the lower bound such that the estimated excess mass to the right of the star threshold ($\sum_{k=\overline{S}}^{s^{ub}} \hat{\omega}_k$) equals the missing mass to the left of the threshold ($\sum_{k=s^{lb}}^{\overline{S}} \hat{\omega}_k$)[65]:

$$\sum_{k=\overline{S}}^{s^{ub}} \hat{\omega}_k = \sum_{k=s^{lb}}^{\overline{S}} \hat{\omega}_k \tag{4}$$

The average bunching response is defined as:

$$\frac{\sum_{k=\overline{S}}^{s^{ub}} \hat{\omega}_k}{0.5(\hat{F}_s^{lb} + \hat{F}_s^{ub})}$$

which represents the average response across all charities, some of whom may not bunch. $\hat{F}_s^{lb}$ is the counterfactual density at the estimated lower bound and $\hat{F}_s^{ub}$ is the counterfactual density at the upper bound of the manipulation region.

Several empirical challenges are worth discussing. The first is that bunching at notches represents intensive-margin responses to incentives, and any extensive-margin response is a threat to identification. Specifically, the concern is that there might be extensive-margin responses below the threshold, such as charity exit, that would shift the distribution down. This would inflate the missing mass relative to the excess mass, meaning that the estimated counterfactual would not be fully stripped of all behavioral responses to the notch. Fortunately, extensive-margin responses are not a

[64] Given that the discontinuity is not quite as stark in my setting, I also try iterating over all possible combinations of s^{lb} and s^{ub}. The results are insensitive to the choice of methodology.

[65] I follow this procedure at each threshold separately.

main issue in the charity sector (compared to the literature on individual taxation), as charities do not exit as a result of the Charity Navigator ratings system[66]. There remains the possibility that charities switch sectors[67], which may appear like exit in the data, but again, this does not seem to be happening[68].

A second, more relevant, empirical challenge is the presence of multiple notches (there are 4 thresholds in the Charity Navigator ratings system). This only poses a problem if bunchers jump over more than one threshold at a time, for example jumping from a 2-star to a 4-star rating. Again, this is because the excess and missing masses would no longer be matched at each notch separately. However, the thresholds are spaced sufficiently far apart to not be an issue here.

A final assumption of the estimation is that bunching only occurs due to charities moving from the left to the right side of the notch. This is violated if there is any bunching from above, which would bias the average bunching response upwards. This may be a concern in this setting if charities well above a threshold have an incentive to slack off. Charity Navigator's rating system encourages charities to score just above 90 - and a further gain does not accrue from increasing the score beyond this level. Charities that score very highly may even decide to increase spending on administration and fundraising, but only to the extent of ensuring a score of at least 90. One way to test this assumption is to examine the behavior of charities that consistently earn 4 stars. If they do indeed slack off and bunch from above, then scores would be expected to fall to 90 over time. Appendix Figure C.5 shows that, on average, the scores of charities that earn 4 stars at least 80% of the time do not decline to 90, or even drop below 93. Ignoring 2011 and 2016 (when the ratings methodology changed), average scores decline from a peak of 97 to 94, but seem to be increasing again in recent years. This suggests that charities are not bunching

[66] In this context, exit can either mean that a charity ceases operations or that it ceases to be rated by Charity Navigator. There is no evidence of either.

[67] Charities are classified according to the National Taxonomy of Exempt Entities (NTEE) classification system, similar to NAICS codes. Thus by switching to a classification code with a more lenient threshold, a charity could potentially improve their rating.

[68] Charities can contact the IRS to try to request a new NTEE classification code, but there are currently no formal procedures in place.

substantially above, and also that there is a limit to the moral hazard problem, as charities are not increasing expenses unnecessarily, even when they face incentives to do so.

5.2.2 Results

Consistent with Figure 2, the bootstrap estimates show that bunching is only significant around the 4-star threshold. Figure 5 depicts the fitted counterfactual alongside binned counts of organizations, and Table 6 presents the results. The sample is restricted to all charities with raw scores over 83, where the threshold is 90 and an upper bound of 92 is used. Pooling all years, Figure 5 shows that, relative to a continuous measure, the 4-star threshold induces charities to increase their overall score. The bunching estimates presented in Table 6 suggest that average scores are 2 points higher than they would be if charities were to be rated on a continuous measure[69]. This result is significant at the 1% level. Furthermore, average scores are over 3 points higher in 2019, and 3.3 points higher in 2015. Bunching can be expected to be more pronounced in these years. Charity Navigator has become more influential over time, and 2015 was the last year financial metrics were derived from current year information (Charity Navigator switched to three-year averaging after June 1st 2016). Charities can increase their score by adjusting several different margins but, by restricting adjustments to administrative expenditure, these estimates imply that, on average, administrative expenses would have to halve in order to achieve this result.

The same methodology can be applied to estimate the behavioral response with respect to Financial Health scores, as well as specific financial metrics. Focusing again on the 4-star threshold, Table 7 presents the bunching estimates for the Financial

[69]Given that the counterfactual distribution is estimated using charities that are located far from the threshold (with little incentive to improve their score), it is likely that it reflects behavior in a world with no score concerns. This is different from charities facing a continuous measure, which would encourage all organizations to improve their score a little.

Health score, and Appendix Table B.7 for the Administrative Expense Percentage[70]. As mentioned in Section 5.1.2, other than the Program Expense Percentage and Program Expense Growth, all other financial metrics are notched. Therefore, in order to estimate the behavioral response with respect to the Financial Health score, I add dummy variables to Equation (2) to control for potential bunching at multiples of 2.5. Pooling all years, and selecting the upper bound score to be 92, Table 7 suggests that average Financial Health scores are 0.5 points higher than if charities were to be rated on a continuous measure. This increases to 2 points higher in 2017 and 2019[71].

As shown in Appendix Figure C.4, charities seem to adjust the Administrative Expense Percentage in particular (recall that a lower ratio is better). Scaling up the Administrative Expense Percentage to be between 0 and 100, and selecting the lower bound as 13, Appendix Table B.7 presents the bunching estimates[72]. The results are mostly insignificant, except in 2018, when charities appear to cut back on the Administrative Expense Percentage by 0.6 percentage points in response to the threshold[73].

5.3 Real changes in behavior versus manipulation

The results thus far have documented that charities are bunching just above the 4-star threshold and that their Financial Health scores are higher than would be expected under a continuous rating system. This section explores whether these responses are the result of charities exerting real effort to change their behavior, or just strategic misreporting. I first consider the cost of manipulation, before presenting two different

[70]Bunching patterns are more diffuse for the other financial metrics.

[71]The reason for the negative coefficient on 2015 may be that, prior to 2016, financial metrics were derived from the most recent Form 990. This means that charities were more able to finetune their overall score, and didn't have to bunch in the Financial Health metric if they didn't need to.

[72]In order for charities to earn full points, they must have an Administrative Expense Percentage less than 15%. This particular sample excludes community foundations, museums and food and humanitarian relief charities, which are rated on a different scale.

[73]Pooling all years, the mean Administrative Expense Percentage is 10.6%.

approaches to uncovering the degree of misreporting - one that looks for traces of evasion, and another that focuses on accounting errors.

5.3.1 Cost of misreporting

Before looking for evidence of misreporting, it is important to first consider whether charities face an incentive to do so. Relative to the case where misreporting is prohibitively costly, evidence of avoidance suggests a lower cost of manipulation. Therefore, before trying to quantify the extent of misreporting, it is useful to consider the cost of avoidance. This is notoriously difficult to measure because the distribution of reported expenses reveals expenditure responses, but not the cost associated with them. However, a common method in the literature (see, for example, Almunia and Lopez-Rodriguez, 2018) is to define a dominated region, where charities do not rationally locate unless resource costs are prohibitively high. The costs borne by those that remain in the dominated region reveal something about the marginal avoidance cost. Comparing the impact of a higher rating on donations to the cost of reducing expenditures helps to define a dominated region and hence the marginal avoidance cost.

One way to understand the costs faced by those in the dominated region is to focus on the cost of reducing specific components of a. For example, what is the effect on donations of reducing fundraising by enough to earn a higher rating? If a charity could earn more donations by cutting fundraising and earning a higher star, then they are classified as being in the dominated region. In practice, this means restricting attention to 3-star charities that are within 2.5 points (out of 100) of the 4-star threshold[74], and then using reported fundraising efficiency ($\frac{fundraising}{donations}$) and $\overline{\delta}$ to calculate the donations they could earn if they were to cut fundraising by enough to earn a 4-star rating. It should be noted that this is an upper bound on the

[74]Charities must be 2.5 points (out of 100) from the threshold, not 1, because financial metrics are also notched, and because the overall score is a weighted sum of both the Financial Health and A&T scores (so improving your Financial Health score by 2.5 points is enough to improve your overall score by 1 point).

dominated region, as fundraising efficiency is likely to be understated for two reasons: first, fundraising is sometimes under-reported, meaning that $\frac{fundraising}{donations}$ will be downward biased; and second, factors other than fundraising help attract donations - some components of administrative spending, and other inputs, are directly related to donations, too.

Over the period of analysis, I find 589 organizations to be within 2.5 points of the 4-star threshold. Of these, 294 are classified as being in the dominated region, and would attract more donations if they were to cut fundraising. Charities in the dominated region report excess fundraising of $100,025, on average. Total expenses in this region are approximately $12 million, which suggests that average resource costs are 1% of total expenses, with the maximum being 7%. In a different setting relating to a firm size threshold in Spain, Almunia and Lopez-Rodriguez (2018) estimate the marginal cost of avoidance as being between 0.06 and 0.19. This suggests that avoidance may be less costly in the nonprofit sector, making misreporting more appealing[75].

5.3.2 Traces of evasion

The first approach to uncovering the degree of misreporting takes inspiration from the "traces of evasion" literature, as summarized by Slemrod and Weber (2012). One incentive that Charity Navigator creates is the incentive to relabel expenses as "program service" expenses. Higher reported spending on program services earns charities a higher score, while reported spending on fundraising or administration reduces their score. It is therefore better to try to classify an administrative expense as a program service expense[76]. In order to try and understand how widespread this behavior is, I compare rated charities with unrated charities for the period 2010-

[75] If the marginal benefits of real and avoidance costs are equated, then we can also think that the costs of changing inputs are low.

[76] Unrated charities also face this incentive if they choose to report expense ratios, however they are unaffected by the rating thresholds and the incentive is therefore a lot weaker.

2017[77].

As with the traces of evasion literature, the idea is to compare metrics expected to be truthfully reported for both rated and unrated charities, to metrics that are likely to be truthfully reported only by unrated charities. Aside from the few organizations that seem to make purposeful accounting errors, I expect all other charities to truthfully report their wage bill and total expenses[78], but only unrated charities to truthfully report program service expenses.

The metrics of interest are thus $\frac{\text{wage bill}}{\text{total expenses}}$ (w_{te}) and $\frac{\text{wage bill}}{\text{program service expenses}}$ (w_{pe}). Step 1 is to calculate these ratios for both rated and unrated charities, and Step 2 is to then determine the relationship between w_{te} and w_{pe} for unrated charities (which we assume to be reporting truthfully). If rated and unrated charities are comparable, then the relationship between these two metrics should be similar for both groups. Indeed, this is equivalent to assuming that true $\frac{\text{program service expenses}}{\text{total expenses}}$ (pe_{te}) is similar for comparable rated and unrated charities. However, if rated charities are relabelling fundraising and administrative expenses as program service expenses, then reported w_{pe} will be understated.

The final step is to then apply the relationship observed among the unrated charities to the rated charities. For example, if w_{pe} is, on average, 1.2 times higher than w_{te} for unrated charities operating in the Arts, Culture and Humanities sector, then I inflate w_{te} of rated charities in the same sector by 1.2 in order to obtain the true ratio.

In order to draw these comparisons, rated and unrated charities need to be similar. Charity Navigator outlines several criteria that must be met in order for a charity to be rated (see Section 2 for details). Therefore, I compare rated charities with charities operating in the same sector meeting the criteria but which, for some reason, are so far unrated. Some organizations, such as hospitals, universities and land trusts, no longer receive ratings, but other charities that meet the criteria are unrated at

[77] This period corresponds to the availability of IRS Form 990 data on e-filers.

[78] The IRS selects returns for review if there are discrepancies between information reported by a payor and payee, e.g. on Form 1099 or W2.

random. A representative from Charity Navigator explained that they "work through that list [of unrated charities] in no particular order".

I test this assertion by restricting the sample to those charities that meet the selection criteria but are either never rated or are rated for the first time during my sample period. Appendix Table B.8 shows that information from the Form 990 is largely unpredictive of whether or not a charity is newly rated. In particular, I find that fundraising, contributions, and volunteers are positively correlated with being selected for a rating, and salary expenses are negatively correlated. However, the R-squared is 0.02, suggesting that there are some Form 990 items that predict being rated but that much unexplained variation remains.

To summarize, three assumptions are used to test these traces of evasion. First, total expenses are reported truthfully by all charities. Aside from "accounting errors", documented in Section 5.3.3, this seems realistic as there is no incentive to understate total expenses, merely to change the composition[79]. Indeed, charities that report expenses that seem too low, given their size, may be flagged by the IRS for review.

The second is that, on average, rated and unrated charities have the same pe_{te}. This is a stronger assumption, although note that I compare charities operating in the same sector, and meeting the same criteria with regard to age[80], size, reliance on public support and fundraising expenses. Appendix Figure C.6 plots the kernel density function of pe_{te} for rated versus unrated charities. It should be noted that this pertains to the reported (and not necessarily true) ratio, but it is nevertheless instructive, and suggests that rated and unrated charities are not too dissimilar in this respect.

The third assumption is that charities that are really exerting effort to cut fundraising and administrative costs do not immediately reallocate the savings to program

[79]Note, if this assumption does not hold, then merely the interpretation of the results changes, not the validity. In particular, misreporting is understated if unrated charities do not report total expenses truthfully, but we can still conclude that they report more truthfully than rated charities.

[80]Rated charities have been in operation for 23 years, on average; unrated charities for 20 years.

expenses. In other words, a cut in fundraising costs translates to a reduction in total expenses and not an increase in program expenses. The argument here is that charities that genuinely cut costs make decisions for spending these cost savings with a lag, whereas charities that are misreporting are relabelling expenses in a given year. If this were not the case, then charities that are exerting genuine effort to cut costs would appear the same as charities misreporting (both would report higher program service expenses). This assumption is difficult to verify, and if some cost savings are immediately reallocated to program service expenses, then I would understate the degree of misreporting.

Appendix Figure C.7 plots the distributions of w_{te} and w_{pe} for rated and unrated charities. It shows that, on the whole, unrated charities spend more on wages as a proportion of total and program expenses[81]. This is consistent with the idea that rated charities face greater pressure to keep costs down.

After following steps 1 and 2, I find that rated charities appear to be overstating program service expenses in several sectors. I calculate both the proportion of rated charities for which w_{pe} is less than the mean ratio they "should" be at, and the proportion that are one standard deviation below this "correct" mean. I find that around 19% of rated charities report w_{pe} to be less than one standard deviation below the "correct" mean[82]. This proportion rises to over 20% in sectors such as Human Services and Community Improvement, and falls to less than 14% in sectors such as Recreation and Medical Research. Furthermore, 75% of organizations in the Food sector report w_{pe} to be less than the "correct" mean. This suggests that overstating program expenses by just a small amount is fairly widespread in this highly competitive sector.

The heterogeneity in misreporting across sectors is as expected, and consistent with the bunching patterns. Administrative expenses are much harder to relabel as pro-

[81]This is particularly noticeable in the food and housing sectors, which are some of the most competitive sectors due to the homogeneous nature of the services they provide.

[82]If w_{pe} followed a normal distribution, then we would expect 16% of observations to fall less than one standard deviation below the mean. However, these are negatively skewed distributions, so fewer than 16% of observations should lie in the left-hand tail.

gram service expenses in health and recreation sectors versus human services and community sectors. For example, office supplies could plausibly be labelled as a program expense if a charity is providing child day care, but not if it is a baseball league. Furthermore, donors who support health-related charities are often loyal to a particular cause, regardless of rating. This gives these charities less of an incentive to misreport by inflating program service expenses.

Finally, I examine how the two metrics evolve after a charity is rated for the first time. If rated charities inflate reported spending on program services, then a decrease in w_{pe} in the years following an initial rating might be expected. The effect on w_{te} is less clear, as some charities' efforts to cut back on administration or fundraising could translate into a real reduction in total expenses, while others that are merely relabelling expenses may leave total expenses unchanged.

I follow the evolution of these metrics for newly rated charities by using a generalized event-study design. This allows for a flexible time path of responses and relies on the assumptions that the outcomes of unrated and newly rated charities would evolve similarly in the absence of rating, and that there are no contemporaneous changes that affect these two groups differentially. Restricting attention to charities that meet the criteria for rating and are rated for the first time during the period of interest, the regressions take the following form:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{s \in \mathcal{S}} \beta_s R_{it} \mathbb{1}\{t = s\} + \gamma R_{it} \mathbb{1}\{t \neq s\} + \epsilon_{it} \tag{5}$$

where Y_{it} is the outcome for charity i at time t and β is the effect of being newly rated on charity i, s periods after the initial rating. R_{it} is a dummy variable equal to one if charity i is rated at time t[83], and $\mathbb{1}\{t = s\}$ is an indicator equal to one if t is s periods away from the rating, where $\mathcal{S} = \{-2, -1, 0, 1, 2\}$. The single parameter γ captures the effect of rating in periods outside $\mathcal{S}$. α_i is a charity fixed effect, λ_t is

[83]Note, I restrict attention to charities where being rated is an absorbing state. In other words, I exclude the very few cases where charities move in and out of being rated.

a time fixed effect, and ϵ_{it} is the error term.

The results are reported in Appendix Table B.9. The coefficients are small and imprecisely estimated. This is because, of the charities that are newly rated between 2010 and 2017, most are rated very early on in the period. As a result, there is insufficient variation in the data to confidently detect a significant effect. However, the interpretation of the coefficients when $w_{pe} * 100$ is the dependent variable, for example, is that a charity that is newly rated in period t reduces $w_{pe} * 100$ by 0.2 percentage points in period t and 0.1 in period $t+2$[84]. This is consistent with the idea that newly rated charities inflate reported spending on program expenses, although the estimated effects are small and insignificant.

In sum, I find evidence of charities misreporting their financials in order to earn a higher star rating. This seems to be concentrated in a few sectors, such as Foundations, Food, Housing and Human Services. Charities operating in these sectors are likely to have different motivations for misreporting, although the food and housing sectors are notoriously competitive and offer fairly homogeneous services[85], meaning that charities face stronger incentives to differentiate themselves through their star rating.

However, misreporting does not seem to be widespread, and may not even be nefarious. Specifically, nonprofits can choose how to allocate the costs of activities that both serve their mission and provide fundraising opportunities[86]. For example, if a literacy nonprofit sends out letters to recruit volunteer tutors and seek donations, it can choose how to allocate the cost of the letters between program and fundraising expenses. The fact that allocation can appear somewhat subjective has come under increased scrutiny, yet accounting and consulting firms still advertise joint allocation

[84] Average $w_{pe} * 100$ in the sample is 53 and the standard deviation is 26.

[85] Gayle et al. (2017) identify public housing as a more competitive and homogeneous sector.

[86] Three criteria must be met in order to apply a joint cost allocation methodology: purpose, audience and content. The purpose of the activity must be to carry out a program function or management and general function. The audience for the fundraising activity cannot be selected based on prior donors, and the content condition is met if the joint activity actually supports program or management and general functions.

as a way to improve fundraising ratios and ratings[87]. Therefore, this section shows that the rating system is encouraging rated charities to take more advantage of these gray areas than unrated organizations.

5.3.3 Accounting errors

The second method is a descriptive test that compares charities that report their finances accurately with those that do not[88]. Figure 6 presents the distribution of scores for charities that understate total expenses, overstate total revenue, and accurately report both[89]. I classify a charity as having understated their total expenses if the sum of every expense line item (administrative expenses, wages, fundraising etc) is greater than the total declared on their "bottom line"[90]. Similarly, a charity overstates their revenue if the sum of all revenue sources is less than the reported total[91].

In general, charities that misreport in this way are much larger than the average rated charity, and are more likely to be scoring 4 stars. They are fairly evenly distributed across sectors, but food banks, human services and foundations are slightly over-represented. Charities operating in these sectors also just bunch more than others.

Figure 6 reveals strong bunching at the 4-star threshold among charities that misreport total expenses or revenue, and little evidence of it among those that accurately

[87] See, for example, https://kahnlitwin.com/blogs/mission-matters-blog/is-my-nonprofit-allocating-joint-costs-properly and https://gbq.com/divide-and-conquer-how-joint-cost-allocating-works/.

[88] This will be an underestimate of the degree of misreporting, as it focuses on a very specific type of misreporting.

[89] I'm assuming here that the components are accurately reported and that it is the totals that are misreported. It seems reasonable that an organization would know how much they had spent on an accountant and get the total wrong, rather than accurately know the total, but not how much they had spent on the accountant.

[90] On average, charities that understate their total expenses underreport by $22, with the largest error being over $17,000. Even the largest error amounts to less than 1% of total expenses.

[91] Most revenue errors are equally small, although there are some outliers where misreporting is close to 20% of total revenue.

report[92]. In other words, focusing on charities that misreport in this particular way, an excess mass just to the right of the 4-star threshold is apparent, suggesting that some are achieving a higher star through misreporting. This pattern would not be found if misreporting was merely due to rounding errors. If there were rounding errors, charities would be just as likely to appear on the left as the right side of the threshold.

Restricting attention to charities that misreport in this way, I recalculate their scores using the correct totals taken from summing expense and revenue components. The panels on the left-hand side of Figure 7 plot the reported scores of charities whose scores fall when re-calculated. The panels on the right-hand side plot these re-calculated scores.

The reported scores provide clear evidence of bunching at the 4-star threshold, particularly in more recent years. However, when the scores are recalculated using the correct totals, the excess mass shifts to the "wrong" side of the threshold. This exercise reveals clear evidence of misreporting, to the extent that some charities earning 4 stars receive only 3 when their scores are calculated using correct totals[93]. Estimating how many charities move to the "wrong" side of the threshold suggests that at least 2% of the bunching at the 4-star threshold (2017-2019) is due to charities under-reporting total expenses and over-reporting total revenue.

Lastly, we can examine whether use of a tax preparer is associated with these types of accounting errors. Given that most charities use a tax preparer, it is unsurprising that the pattern observed in the full sample is repeated in Appendix Figure C.8. The sample sizes are a little too small to conclude anything about misreporting among charities that do not use a preparer (Appendix Figure C.9), although the excess mass on the 4-star side of the threshold does disappear when the correct totals are used. Of greater interest, is the difference between experienced and CPA-certified preparers.

[92]We do still see some evidence of bunching at the 3-star threshold among those that accurately report. This may be suggestive of real effort or just that these charities are misreporting in a different way.

[93]There also appears to be some misreporting of perfect scores. Namely, the mass of charities at 100 disappears when scores are recalculated using correct expense and revenue totals.

The former make fewer errors, and the excess mass does not shift to the 3-star side of the threshold when the correct totals are used (Appendix Figure C.10). However, CPA-certified preparers appear to misreport to the extent that some charities lose their 4-star rating when scores are calculated using the correct totals (Appendix Figure C.11).

6 Lessons learned

This final section of the paper uses the preceding analysis to discuss the lessons learned on how to optimally rate charities. The question I am best equipped to answer is how a notched rating system compares to a continuous measure.

As discussed in Section 3, Charity Navigator's mission statement implies that two of its main objectives are to reallocate resources to more efficient charities, and to promote waste reduction. The first objective relates to selection: Charity Navigator wishes to shift resources to more effective charities[94]. By creating a rating system, Charity Navigator provides donors with information relating to which charities minimize costs and have good accountability and governance. The second objective relates to moral hazard - the absence of pressure on managers to minimize costs means that nonprofits are likely to incur greater expenses than their for-profit counterparts. The introduction of a rating system can change incentives such that nonprofits cut back on waste in return for a higher star rating.

Moving from a star rating system to a continuous measure affects the achievement of these objectives. Specifically, the reallocation objective is linked to the donor response, while the waste reduction objective relates to bunching and misreporting[95].

[94] Diminishing returns to inputs means that allocating all donations to just one charity is suboptimal.

[95] It also relates to take-up of information, but I am unable to measure this.

6.1 Reallocation to more effective charities

The first objective of Charity Navigator is reallocation. By providing information on which charities report good governance and financial health, Charity Navigator hopes to shift resources to more effective charities. A notched rating system encourages donors to support charities with 3- or 4-star ratings, but, unlike a continuous measure, does not distinguish between charities with a raw overall score of 90 versus 100. In other words, a notched system is effective at redistribution but is a relatively blunt instrument, with charities not rewarded for increasing their score beyond the star threshold.

Furthermore, ratings systems can create both winners and losers. Unless total donations are increasing, some charities will see donations fall as a result of a rating system. It is hard to identify the charities that would suffer under a continuous measure. However, the results reported in Section 4.1 suggest that some of the extra donations awarded to 3- and 4-star charities are at the expense of 2-star charities.

6.2 Waste reduction and misreporting

The impact of a rating system on waste reduction depends on how much the charity values a higher rating: the more valuable the rating, the stronger the incentive to cut back on waste in order to achieve the higher rating. Comparing the star rating system to a continuous measure, the incentive to reduce waste will be stronger for some charities under the former as the reward for crossing a star threshold is higher than the reward for a one-point increase in the charity's score. However, this is likely to apply only at the thresholds - charities that score 82 points under the notched system have less of an incentive to increase their score by one point than under a continuous measure. The bunching analysis quantifies the precise reduction in expenditures achieved by charities that are close to the threshold, for a notched rating system compared with a continuous measure. In this way, the bunching analysis shows that Charity Navigator moves closer to achieving its objective of waste reduction under

a notched rating system. However, this applies only for charities that are close to the thresholds, and is at the expense of increased deadweight loss, as organizations spend resources in trying to cross the thresholds.

Although the model outlined in Section 3 equates real and avoidance costs, the evidence presented in Section 5.3 suggests that charity responses comprise both real changes in expenditures and misreporting. If the value of a higher rating is higher under a notched system, so is the value of misreporting. This means that charities close to the thresholds face a greater incentive to misreport under the star rating system as compared to the continuous measure[96].

6.3 Evaluation

The question whether a continuous measure is more efficient than a notched rating system is important, not least because of the recent launch of Charity Navigator's new Encompass Rating System. This latest methodology is still being developed in beta, but it moves away from star ratings towards a continuous measure, with organizations scored between 0 and 100. Furthermore, it expands eligibility to all charities that have e-filed the IRS Form 990[97] for three consecutive years. This increases the number of rated charities from 9,000 to over 160,000, with smaller, less established organizations no longer disqualified from being rated. The preceding analysis can therefore provide information about the potential impacts of the new system, in addition to informing policymakers on the optimal method of charity rating.

However, some questions remain unanswered, and are beyond the scope of the paper. Specifically, the focus is on local analysis, so any questions related to the overall impact of Charity Navigator cannot be answered. Throughout the paper I take

[96]Considering the value of misreporting in an Allingham-Sandmo framework (1972), the probability of detection and the penalty rate remain fixed across two ratings systems, but the gain from misreporting changes. Specifically, the gain from misreporting is higher for charities close to the thresholds under a notched system versus a continuous measure.

[97]Forms 990PF, 990EZ or 990N are not sufficient.

Charity Navigator's objective function as given. A discussion of its wider influence would need to consider alternative objectives, which I save for future work.

The other limitation is a lack of information on donors. Without this, it is impossible to know how much weight donors place on salience of information versus search costs, for example. If salience is important, then synthesizing information under a notched rating system can be welfare-improving if the cost of understanding a continuous measure is as high as finding the information oneself. If, however, search costs are the biggest obstacle, then providing information on more charities (as the Encompass Rating System does) is of greater utility to donors.

The preceding analysis suggests that a notched rating system helps to overcome the information problem faced by donors, and encourages charities to cut back on wasteful expenditures. However, the loss to society is that it also provides charities close to the thresholds with the opportunity for obfuscation[98]. The relatively low cost of avoidance suggests that the practice could be widespread.

It is only possible to comprehensively investigate whether the Encompass Rating System is an improvement on the Star Rating System once it launches, but the analysis contained in this paper suggests that a notched rating system induces greater behavioral change than a continuous measure, although affects a smaller number of charities. This trade-off between the size of the impact and the number of affected agents is not unique to this setting. It appears throughout the tax system, from the design of the Earned Income Tax Credit (Liebman, 2002) to subsidies for fuel-efficient cars (Sallee and Slemrod, 2009). In general, the welfare cost, or gain, of a policy that features a notch or a kink depends on the available alternatives relative to the second-best optimum that is unconstrained by functional form (Slemrod, 2013).

[98] That is not to say that charities would not misreport under a continuous measure - the incentives would just be weaker.

7 Conclusion

This paper investigates the effects of the rating system used by Charity Navigator, the largest evaluator of charities in the United States. I find that donors reward charities for a higher star rating such that a one-star increase in rating from 3 to (the highest possible) 4 stars results in a 6% rise in contributions the following period. This effect is even larger for smaller charities, which experience an 11% rise in contributions. However the impact is not symmetric - earning a lower star does not significantly reduce donations.

Charities respond to these incentives by bunching just over the star thresholds, especially the highest 4-star threshold. Charities bunching at the 4-star threshold score over 2 points higher than if they were to be rated on a continuous measure. This is mainly due to charities scoring better on Financial Health metrics, and in particular, reporting spending less on administration and fundraising. However, evidence of misreporting suggests that the charity response is not entirely real. In particular, some charities appear to relabel expenses as program service expenses, while others overstate their revenue and understate expenses. When scores are recalculated using the correct totals, some of these charities find themselves on the "wrong" side of the threshold, suggesting that the rating system also has the unintended consequence of inducing some charities to misreport.

Based on these findings, it is clear that a star rating system induces greater change for charities located close to the thresholds, whereas a continuous measure would have a smaller effect, but on a larger number of organizations. The design of an optimal rating system depends both on the relative value of these two effects and on donor preferences. If salience is important, then synthesizing information under a notched rating system can be welfare-improving. This is true not only in the charity context, but in many other settings, such as the rating of schools, healthcare services and consumer products. This paper highlights that optimal rating design must take into account both sides of the market.

8 References

Allingham, M. and Sandmo, A. (1972) "Income Tax Evasion: A Theoretical Analysis," Journal of Public Economics, 1(3-4): 323?338

Almunia, M. and Lopez-Rodriguez, D. (2018) "Under the Radar: The Effects of Monitoring Firms and Tax Compliance," American Economic Journal: Economic Policy 10(1): 1-38

Brown, A., Meer, J. and Williams, J. (2017) "Social Distance and Quality Ratings in Charity Choice," Journal of Behavioral and Experimental Economics 66: 9-15

Calabrese, T. (2011) "Do Donors Penalize Nonprofit Organizations with Accumulated Wealth?," Public Administration Review 71(6): 859-869

Calabrese, T. and Grizzle, C. (2012) "Debt, Donors, and the Decision to Give," Journal of Public Budgeting, Accounting & Financial Management 24(2): 221-254

Cattaneo, M., Jansson, M. and Ma, X. (2018) "Manipulation Testing Based on Density Discontinuity," The Stata Journal 18(1): 234-261

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

Chetty, R., Friedman, J., Olsen, T. and Pistaferri, L. (2011) "Adjustment Costs, Firm Responses, and Micro vs Macro Labor Supply Elasticities: Evidence From Danish Tax Records," Quarterly Journal of Economics 126: 749-804

Darden, M. and McCarthy, I. (2015) "The Star Treatment: Estimating the Impact of Star Ratings on Medicare Advantage Enrollments," The Journal of Human Resources 50(4): 980-1008

Deryugina, T. and Marx, B. (2021) "Is the Supply of Charitable Donations Fixed? Evidence from Deadly Tornadoes," American Economic Review: Insights, forthcoming

Dranove, D. and Jin, G. (2010) "Quality Disclosure and Certification: Theory and Practice," Journal of Economic Literature 48(4): 935-963

Fang, L. (2019) "The Effects of Online Review Platforms on Restaurant Revenue, Survival Rate, Consumer Learning and Welfare," Working Paper

Gayle, P., Harrison, T. and Thornton, J. (2017) "Entry, donor market size, and competitive conduct among nonprofit firms," International Journal of Industrial Organization 50: 294-318

Glaeser, E. and Shleifer, A. (2001) "Not-for-Profit Entrepreneurs," Journal of Public Economics 81(1): 99-115

Harris, E. and Neely, D. (2016) "Multiple Information Signals in the Market for Charitable Donations," Journal of Accounting, Auditing & Finance 36(1): 195-220

Harris, E. and Neely, D. (2021) "Determinants and Consequences of Nonprofit Transparency," Contemporary Accounting Research 33(3): 989-1012

Homonoff, T., Spreen, T. and St. Clair, T. (2020) "Balance Sheet Insolvency and Contribution Revenue in Public Charities," Journal of Public Economics 186: 104177

Jin, G. (2005) "Competition and Disclosure Incentives: An Empirical Study of HMOs," RAND Journal of Economics 118(3): 843-877

Kleven, H. and Waseem, M. (2013) "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan," Quarterly Journal of Economics 128(2): 669-723

Krasteva, S. and Yildirim, H. (2013) "(Un)infomed Charitable Giving," Journal of Public Economics 106: 14-26

Lee, D. and Lemieux, T. (2010) "Regression Discontinuity Designs in Economics," Journal of Economic Literature 48(2): 281-355

Leuz, C., Triantis, A. and Wang, T. (2008) "Why Do Firms Go Dark? Causes and

Economic Consequences of Voluntary SEC Deregulations," Journal of Accounting and Economics 45(2-3): 181-208

Lewis, G and Zervas, G. (2016) "The Welfare Impact of Consumer Reviews: A Case Study of the Hotel Industry," Working Paper

Liebman, J. (2002) "The Optimal Design of the Earned Income Tax Credit," in Bruce D. Meyer and Douglas Holtz-Eakin (eds.), Making Work Pay: The Earned Income Tax Credit and Its Impact on American Families. New York, NY, Russell Sage Foundation

Luca, M. (2016) "Reviews, Reputation, and Revenue: The Case of Yelp.com," Harvard Business School NOM Unit Working Paper No. 12-016

Marx, B. (2018) "The Cost of Requiring Charities to Report Financial Information," Working Paper

Mayo, J. (2021) "How Do Big Gifts Affect Rival Charities and Their Donors," Journal of Economic Behavior and Organization 191: 575-597

Okten, C. and Weisbrod, B. (2000) "Determinants of Donations in Private Nonprofit Markets," Journal of Public Economics 75(2): 255-272

Reimers, I. and Waldfogel, J. (2020) "Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings," NBER Working Paper No. 26776

Saez, E. (2010) "Do Taxpayers Bunch at Kink Points?," American Economic Journal: Economic Policy 2(3): 180-212

Sallee, J. and Slemrod, J. (2012) "Car Notches: Strategic Automaker Responses to Fuel Economy Policy," Journal of Public Economics 96: 981-999

Slemrod, J. (2013) "Buenas Notches: Lines and Notches in Tax System Design," eJournal of Tax Research 11(3): 259-283

Slemrod, J. and Weber, C. (2012) "Evidence of the Invisible: Toward a Credibility

Revolution in the Empirical Analysis of Tax Evasion and the Informal Economy," International Tax and Public Finance 19(1): 25-53

St. Clair, T. (2016) "How Do Non-Profits Respond to Regulatory Thresholds: Evidence From New York's Audit Requirements," Journal of Policy Analysis and Management 35(4): 772-790

Velayudhan, T. (2019) "Misallocation or Misreporting? Evidence from a Value Added Tax Notch in India," Working Paper

Yörük, B. (2012) "The Effect of Media on Charitable Giving and Volunteering: Evidence from the "Give Five" Campaign," Journal of Policy Analysis and Management 31(4): 813-836

Yörük, B. (2016) "Charity ratings," Journal of Economics and Management Strategy 25: 195-219

9 Tables

Table 1: Summary Statistics for Rated Charities

	Mean (Std Dev)
Raw score	85.0
	(9.2)
Financial Health score	84.7
	(10.8)
Accountability & Transparency score	89.8
	(10.0)
Net assets	27,786,964
	(159,067,248)
Revenue	14,813,510
	(71,551,484)
Contributions	11,183,576
	(49,440,270)
Total expenses	13,604,523
	(65,899,212)
Program expenses	11,538,000
	(58,714,457)
Administrative expenses	1,137,551
	(5,376,503)
Fundraising expenses	924,866
	(4,863,058)
Observations	148133

Notes: All dollars are constant ($2010)

Standard deviations are in parentheses

Table 2: Effect on Donations of an Increase in Rating From 3⋆ to 4⋆

	All charities	Small	Medium	Large
Rating increased	0.0571***	0.113***	0.0474***	0.0880*
	(0.0145)	(0.0367)	(0.0152)	(0.0532)
Fundraising expenses ($ million)	0.0116**	0.575***	0.0739***	0.00720
	(0.00582)	(0.141)	(0.0267)	(0.00515)
Net assets ($ million)	-0.000343*	-0.0102	-0.00981***	-0.000162
	(0.000181)	(0.0189)	(0.00221)	(0.000156)
Fundraising efficiency	-0.431***	-0.755**	-0.463***	-0.776**
	(0.108)	(0.305)	(0.121)	(0.316)
Charity Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Observations	16272	1628	13017	1627
R-Squared	.0505	.1080	.0620	.0496

Notes: The dependent variable is log(contributions)

All explanatory variables are lagged by one period

The sample is restricted to charities with a 3-star rating in 2015

Small charities are those with net assets less than $1 million (10th percentile)

Large charities are those with net assets greater than $57 million (90th percentile)

Standard errors (in parentheses) clustered at the charity level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Effect on Donations of an Increase in Rating From 2⋆ to 3⋆

	All charities	Small	Medium	Large
Rating increased	0.0768***	0.0765	0.0765***	-0.00439
	(0.0193)	(0.0640)	(0.0216)	(0.0659)
Fundraising expenses ($ million)	0.0125	0.350	0.115***	0.00317
	(0.00984)	(0.308)	(0.0336)	(0.00483)
Net assets ($ million)	0.00228*	0.0182	-0.00379	0.00247*
	(0.00130)	(0.0403)	(0.00681)	(0.00142)
Fundraising efficiency	-0.402***	-1.178***	-0.405***	-0.376
	(0.111)	(0.392)	(0.109)	(0.312)
Charity Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Observations	6389	638	5113	638
R-Squared	.0657	.1717	.0632	.1765

Notes: The dependent variable is log(contributions)

All explanatory variables are lagged by one period

The sample is restricted to charities with a 2-star rating in 2015

Small charities are those with net assets less than $354,000 (corresponding to the 10th percentile)

Large charities are those with net assets greater than $23 million (corresponding to the 90th percentile)

Standard errors (in parentheses) clustered at the charity level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Effect on Donations of a Decrease in Rating From 4⋆ to 3⋆

	All charities	Small	Medium	Large
Rating decreased	-0.00231	-0.00804	-0.00371	-0.162
	(0.0264)	(0.0365)	(0.0284)	(0.102)
Fundraising expenses ($ million)	0.0178***	0.379**	0.0938***	0.00851*
	(0.00573)	(0.172)	(0.0218)	(0.00448)
Net assets ($ million)	-0.0000971*	0.00542	-0.00936***	-0.0000715*
	(0.0000508)	(0.0212)	(0.00232)	(0.0000392)
Fundraising efficiency	-0.496***	-1.272***	-0.683***	0.943
	(0.184)	(0.473)	(0.205)	(0.838)
Charity Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Observations	10951	1095	8847	1095
R-Squared	.0696	.2315	.0712	.0985

Notes: The dependent variable is log(contributions)

All explanatory variables are lagged by one period

The sample is restricted to charities with a 4-star rating in 2015

Small charities are those with net assets less than $1.8 million (10th percentile)

Large charities are those with net assets greater than $73 million (90th percentile)

Standard errors (in parentheses) clustered at the charity level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Effect on Donations of a Decrease in Rating From 3⋆ to 2⋆

	All charities	Small	Medium	Large
Rating decreased	-0.0205	-0.0232	-0.0211	0.0537
	(0.0345)	(0.0493)	(0.0500)	(0.123)
Fundraising expenses ($ million)	0.0109**	0.515***	0.0697***	0.00637
	(0.00540)	(0.133)	(0.0230)	(0.00503)
Net assets ($ million)	-0.000399**	-0.0146	-0.0107***	-0.000221
	(0.000186)	(0.0120)	(0.00253)	(0.000161)
Fundraising efficiency	-0.300**	-0.749***	-0.372***	-0.687*
	(0.122)	(0.279)	(0.138)	(0.404)
Charity Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Observations	12955	1295	10481	1295
R-Squared	.0437	.1188	.0557	.0584

Notes: The dependent variable is log(contributions)

All explanatory variables are lagged by one period

The sample is restricted to charities with a 3-star rating in 2015

Small charities are those with net assets less than $863,000 (corresponding to the 10th percentile)

Large charities are those with net assets greater than $52 million (corresponding to the 90th percentile)

Standard errors (in parentheses) clustered at the charity level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Overall score - 4 $\star$ threshold

	Average bunching response	Observations
All Years	1.849***	58657
	(0.335)	
2019	3.146***	7785
	(0.490)	
2018	2.344***	8222
	(0.421)	
2017	2.800***	6356
	(0.425)	
2015	3.273***	6518
	(0.491)	

Notes: Bootstrapped standard errors are shown in parentheses

Sample restricted to charities that score 83 points or more

Upper bound is selected as 92

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Financial Health score - 4 $\star$ threshold

	Average bunching response	Observations
All Years	0.518***	51962
	(0.127)	
2019	2.113***	5420
	(0.648)	
2018	1.478***	7044
	(0.546)	
2017	2.072***	5738
	(0.689)	
2015	-1.527***	5710
	(0.503)	

Notes: Bootstrapped standard errors are shown in parentheses

Upper bound is selected as 92

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

10 Figures

Figure 1: Search results for "Detroit Institute of Arts"

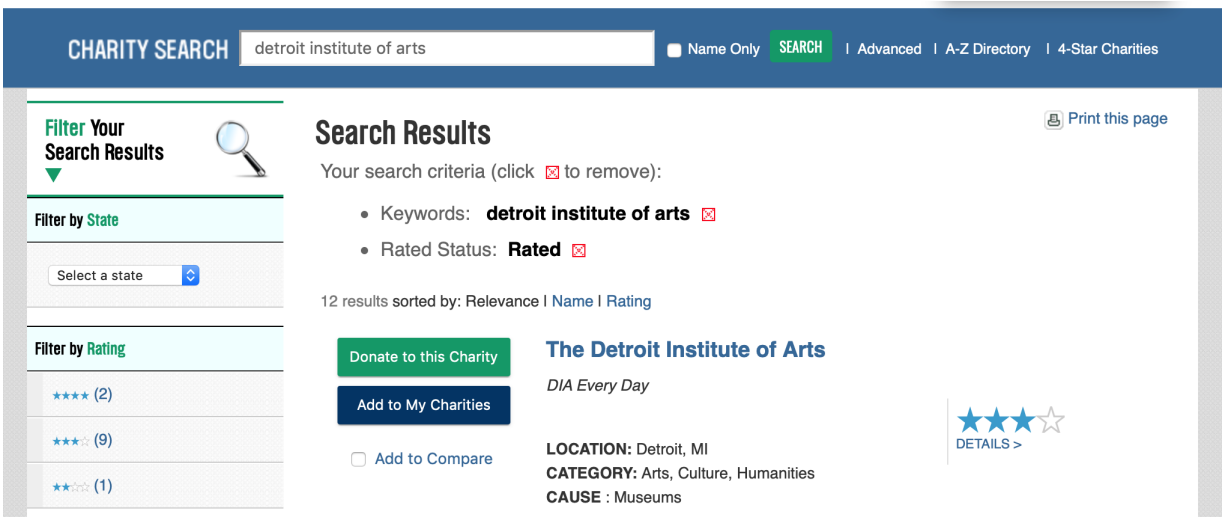


Figure 2: Bunching in the Overall Score - 2002-19

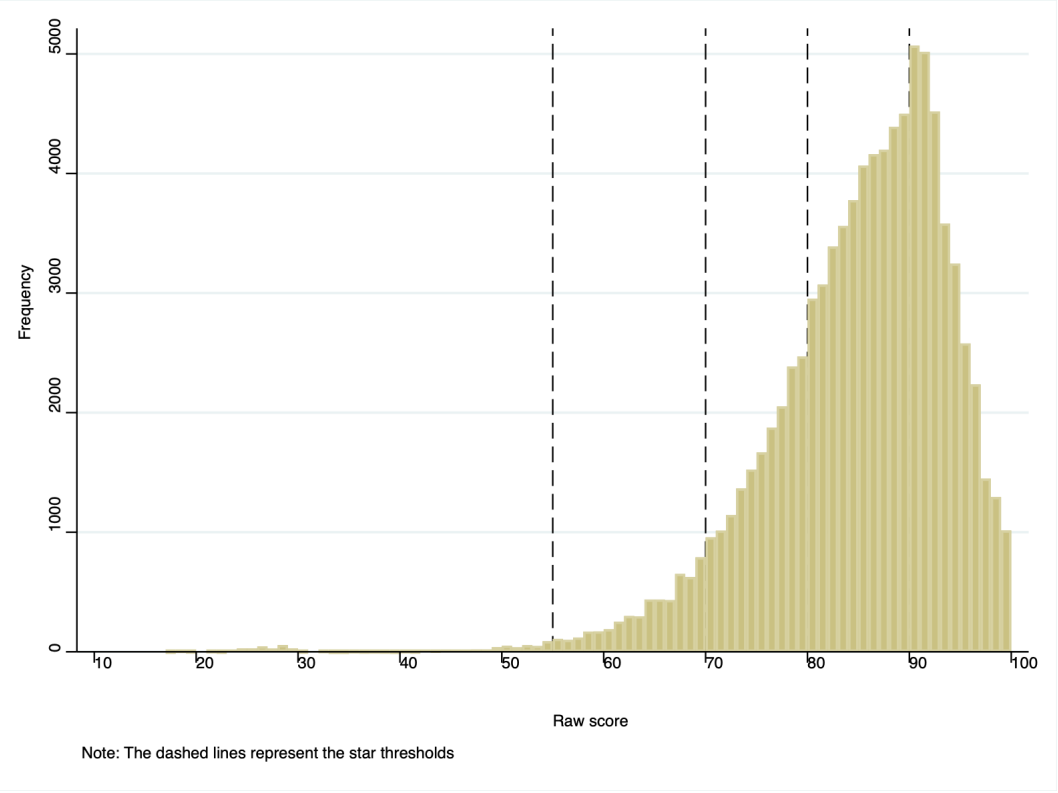


Figure 3: Bunching in the Overall Score - 2016-19

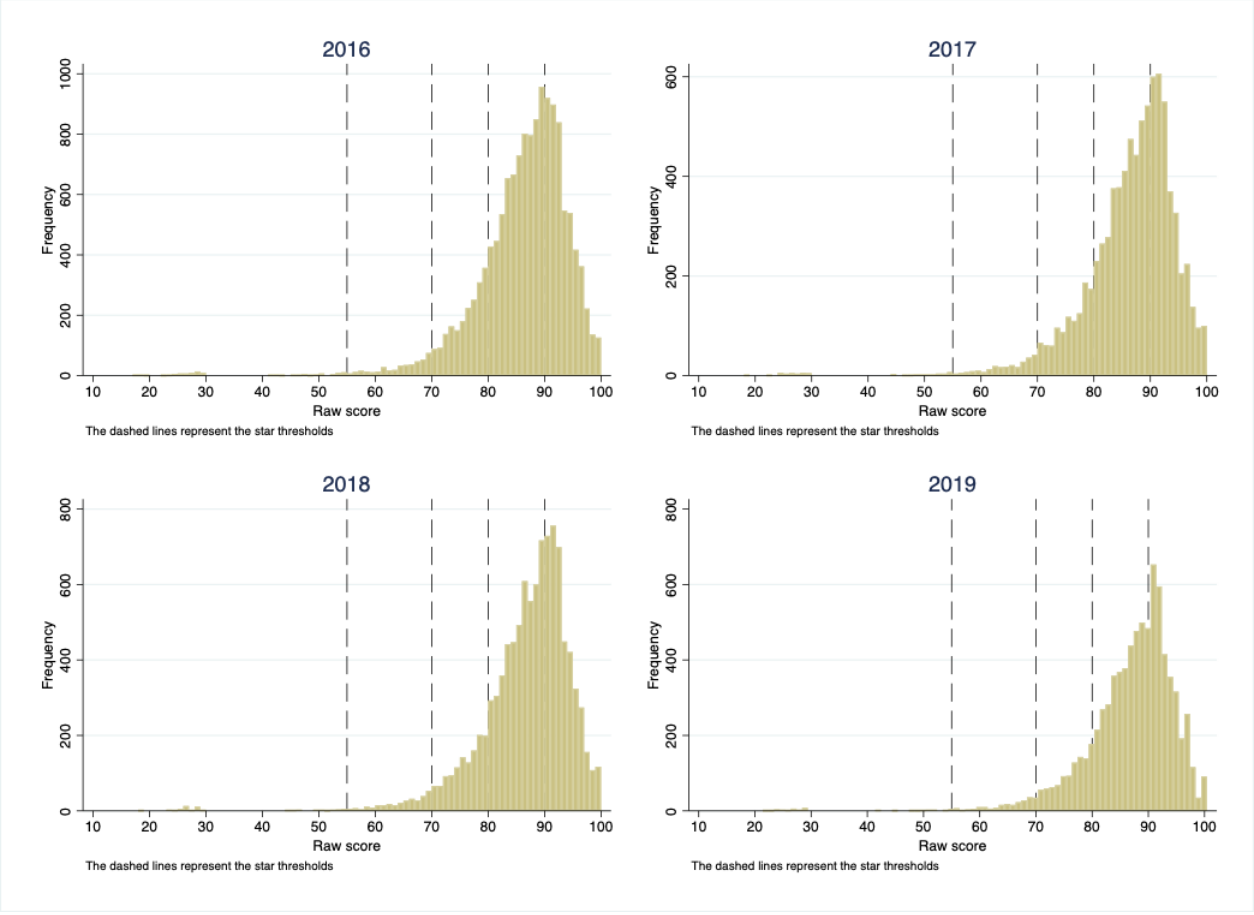


Figure 4: Bunching in the Financial Health and A&T Scores

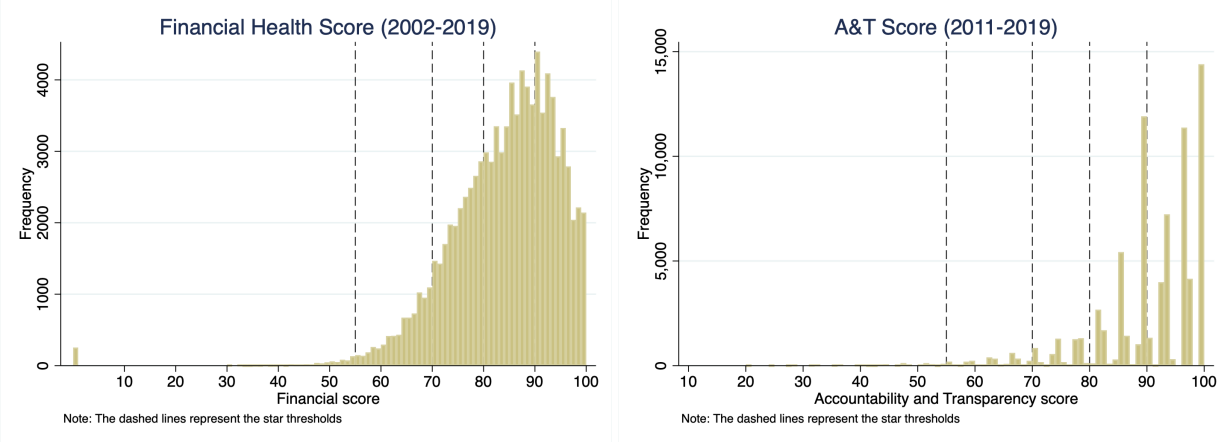


Figure 5: Bunching at the 4-star threshold - 2002-19

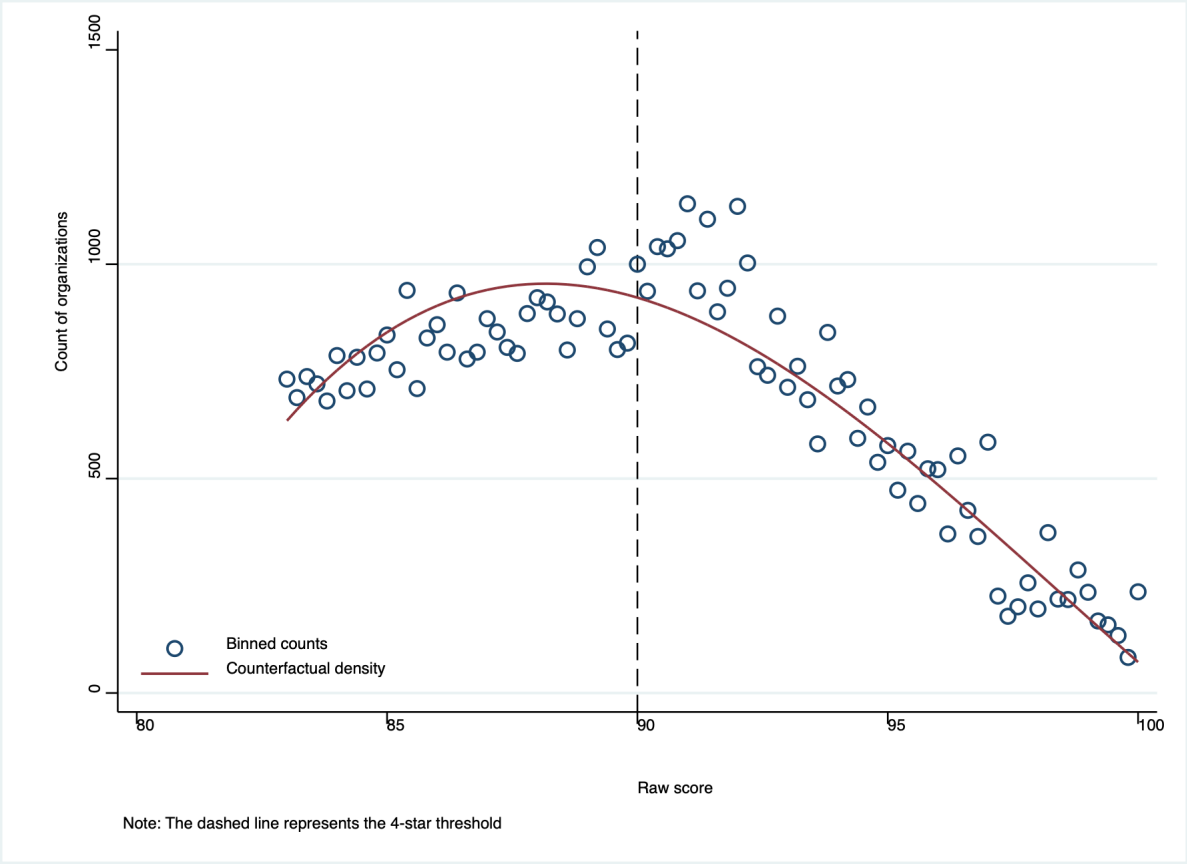


Figure 6: Misreporting

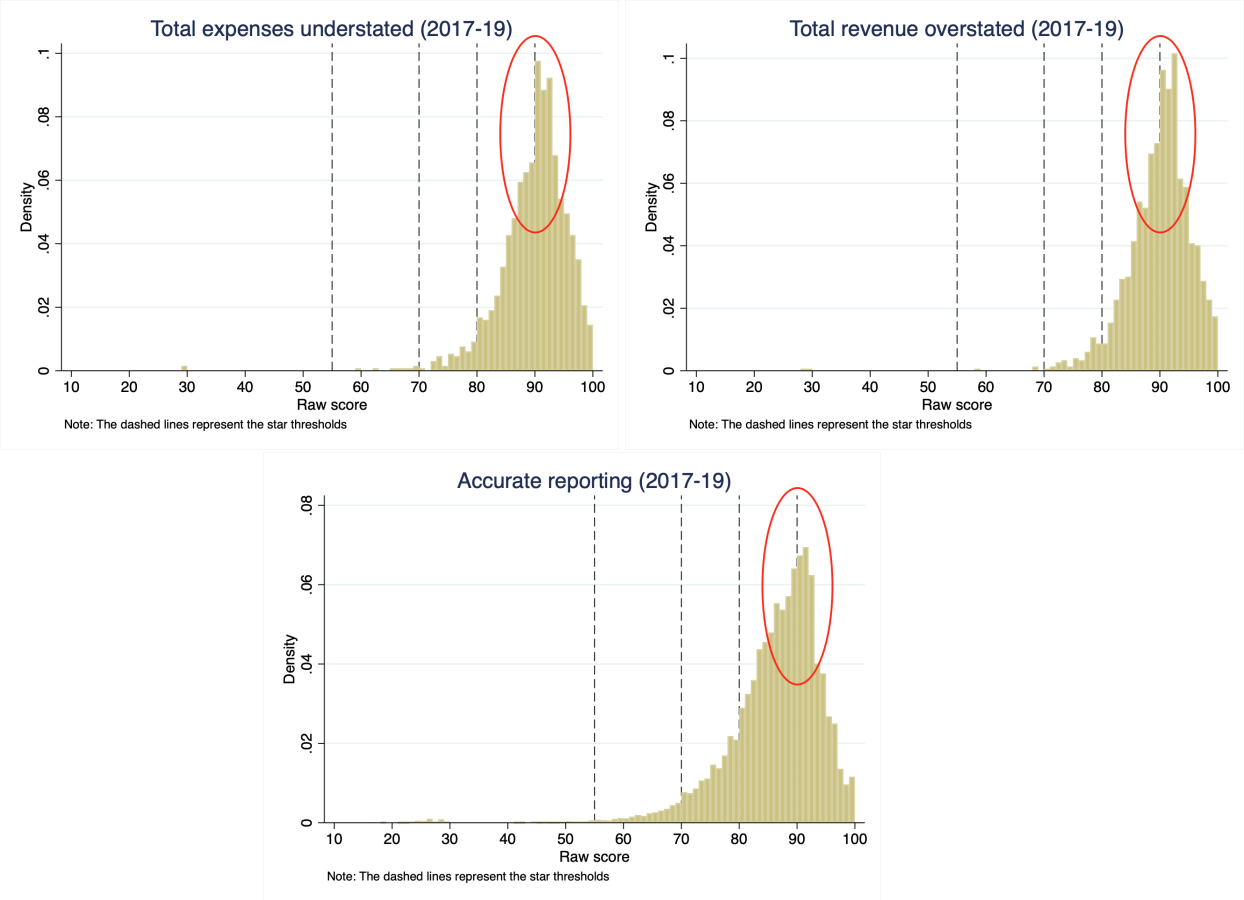
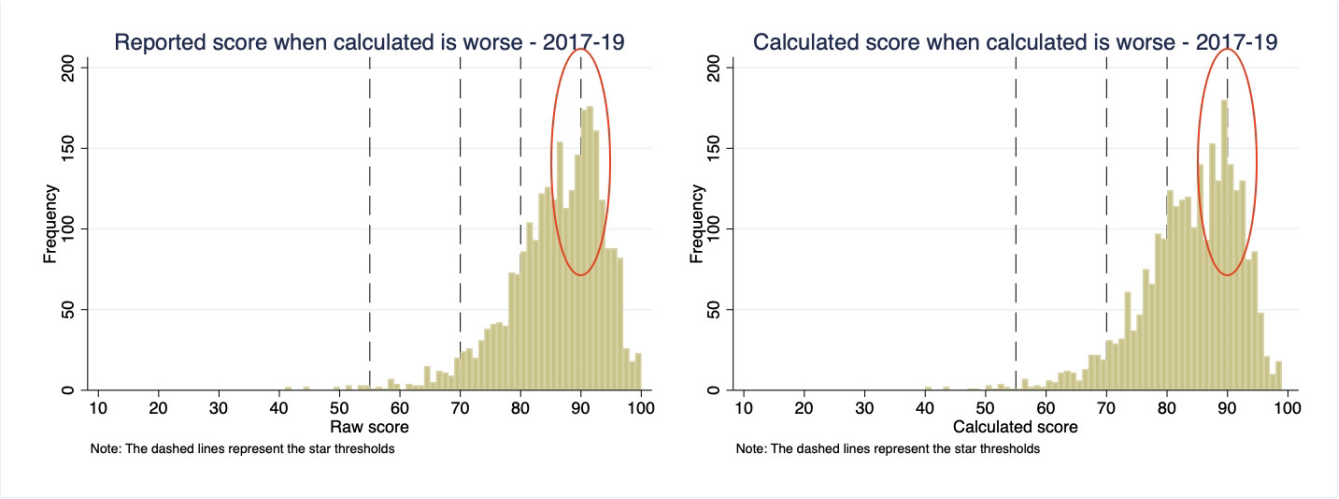


Figure 7: Corrected Scores



A Real and avoidance costs

In order to demonstrate the equivalence between the marginal benefits of real and avoidance costs, I consider organizations that use only one input, a. Returns to scale are denoted by $\omega \in [0, 1)$ and, assuming a quadratic cost function, the charity's problem can be written as:

$$\max_{a,m} \{g(a, D, O) - \frac{(a - \hat{a})^2}{\psi}\}$$

Charities choose a and m[99] to maximize the production of charitable services, such that the first order conditions are as follows:

$$g_a(a, D, O) + g_D(a, D, O)(D_s \cdot s_a) - \frac{2}{\psi}(a - \hat{a}) = 0$$

$$g(a, D, O)(D_s \cdot s_m) - \frac{2}{\psi}(a - \hat{a}) = 0$$

[99]Recall that $m := (a - \hat{a}) \geq 0$, where $\hat{a}$ is reported inputs.

As described in Section 3, the marginal benefit of actually increasing expenditure on a comes from the increase in the production of charitable services. However, this is weighed against the increased cost of inputs, as well as the negative effect on donations (via the score). The marginal benefit of increasing misreporting comes from increased donations. This is weighed against the cost of misreporting. The important thing to note here is that $C_a(a, m, \psi) = \frac{2}{\psi}(a - \hat{a}) = C_u(a, m, \psi)$. In other words, since there are no externalities in the model, the welfare cost of real changes in input is equivalent to the welfare cost of misreporting. In practice, this implies that hiring an accountant to help you cut costs generates the same amount of deadweight loss as misreporting.

B Appendix Tables

Table B.1: Accountability and Transparency Metrics

Performance Metric	Deductions from Score
Less than 5 independent voting members of the board	15 points
Independent members are not a voting majority	15 points
Material diversion of assets in last 2 years, without explanation	15 points
Financial statements not prepared/reviewed by independent accountant	15 points
Material diversion of assets in last year, with explanation	7 points
Independent accountant not selected or overseen by internal committee	7 points
Loans to or from officers or other interested parties	4 points
Organization does not keep board meeting minutes	4 points
Forms 990 not distributed to the board before filing	4 points
No Conflict of Interest policy	4 points
No Whistleblower policy	4 points
No Records retention and destruction policy	4 points
Does not properly report CEO compensation on Form 990	4 points
No process for reviewing and updating CEO compensation	4 points
Fails to report board members and compensation fully on the Form 990	4 points
Does not publish board members on website	4 points
Does not publish latest Audited Financial Statements on website	4 points
No donor privacy policy	4 points
Does not publish senior staff on website	3 points
Does not publish latest Form 990 on website	3 points
Opt-out donor privacy policy	3 points

Table B.2: NTEE Classification of Rated Charities

	Rated charities
A: Arts, Culture and Humanities	1291
B: Education	483
C: Environment	395
D: Animal-Related	444
E: Health Care	387
F: Mental Health and Crisis Intervention	103
G: Medicine and Voluntary Health Associations	286
H: Medical Research	137
I: Crime and Legal-Related	191
J: Employment	62
K: Food, Agriculture and Nutrition	300
L: Housing and Shelter	321
M: Public Safety, Disaster Preparedness and Relief	28
N: Recreation and Sports	222
O: Youth Development	467
P: Human Services	1135
Q: International, Foreign Affairs and National Security	592
R: Civil Rights, Social Action and Advocacy	178
S: Community Improvement and Capacity Building	169
T: Philanthropy, Voluntarism and Grantmaking	733
U: Science and Technology Research Institutes	58
V: Social Science Research Institutes	43
W: Public and Society Benefit	141
X: Religion-Related	463
Y: Mutual and Membership Benefit	6
Total	8635

Table B.3: Effect of Star vs Score

	Effect of Star	Effect of Score
Star rating	0.0760***	
	(0.00443)	
Overall score		0.00846***
		(0.000509)
Fundraising expenses ($ million)	0.0195***	0.0192***
	(0.00589)	(0.00588)
Net assets ($ million)	0.0000113	0.00000775
	(0.0000946)	(0.0000957)
Fundraising efficiency	-0.0800	-0.0620
	(0.0827)	(0.0697)
Charity Fixed Effects	Yes	Yes
Time Fixed Effects	Yes	Yes
Observations	78813	70943
R-Squared	.2430	.2441

Notes: The dependent variable is log(contributions)
All explanatory variables are lagged by one period
Scores that are 1 point away from a threshold are omitted
Standard errors (in parentheses) clustered at the charity level
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.4: Effect on Donations of an Increase in Rating From 3⋆ to 4⋆ - 2015-17

	All charities	Small	Medium	Large
Rating increased	0.0497***	0.0998***	0.0390***	0.101
	(0.0144)	(0.0342)	(0.0145)	(0.0669)
Fundraising expenses ($ million)	0.00987	0.281**	0.0439**	0.00606
	(0.00843)	(0.127)	(0.0188)	(0.00706)
Net assets ($ million)	-0.000181	-0.0283	-0.0180***	0.000140
	(0.000185)	(0.0235)	(0.00304)	(0.000127)
Fundraising efficiency	-0.363***	-0.698**	-0.463***	-0.537**
	(0.118)	(0.292)	(0.134)	(0.272)
Charity Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Observations	12625	1093	10398	1134
R-Squared	.0354	.0896	.0557	.04634

Notes: The dependent variable is log(contributions)

All explanatory variables are lagged by one period

The sample is restricted to charities with a 3-star rating in 2015

Small charities are those with net assets less than $1 million (10th percentile)

Large charities are those with net assets greater than $57 million (90th percentile)

Standard errors (in parentheses) clustered at the charity level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.5: Effect on Donations of an Increase in Rating From 3⋆ to 4⋆ - IV reduced form using 3-year averaging

	All charities	Small	Medium	Large
Rating increased	0.0166	0.0943	0.00163	0.0694
	(0.0257)	(0.0741)	(0.0244)	(0.0998)
Fundraising expenses ($ million)	0.0113*	0.576***	0.0734***	0.00668
	(0.00580)	(0.141)	(0.0268)	(0.00508)
Net assets ($ million)	-0.000340*	-0.0100	-0.00994***	-0.000167
	(0.000181)	(0.0192)	(0.00222)	(0.000156)
Fundraising efficiency	-0.427***	-0.800***	-0.460***	-0.772**
	(0.109)	(0.302)	(0.121)	(0.319)
Charity Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Observations	16272	1628	13017	1627
R-Squared	.0485	.1014	.0605	.0469

Notes: The dependent variable is log(contributions)

All explanatory variables are lagged by one period

The sample is restricted to charities with a 3-star rating in 2015

Charities that would have experienced a rating increase under 3-year averaging are used as instruments for those whose rating actually increased

Small charities are those with net assets less than $1 million (10th percentile)

Large charities are those with net assets greater than $57 million (90th percentile)

Standard errors (in parentheses) clustered at the charity level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.6: Effect on Donations of an Increase in Rating From 3⋆ to 4⋆ - IV reduced form fixing scores

	All charities	Small	Medium	Large
Rating increased	0.119***	0.147**	0.107***	0.153***
	(0.0211)	(0.0649)	(0.0225)	(0.0577)
Fundraising expenses ($ million)	0.0116**	0.582***	0.0745***	0.00718
	(0.00583)	(0.142)	(0.0269)	(0.00512)
Net assets ($ million)	-0.000348*	-0.0108	-0.00991***	-0.000170
	(0.000181)	(0.0192)	(0.00222)	(0.000156)
Fundraising efficiency	-0.429***	-0.814***	-0.463***	-0.763**
	(0.109)	(0.304)	(0.121)	(0.321)
Charity Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Observations	16272	1628	13017	1627
R-Squared	.0527	.1032	.0644	.0516

Notes: The dependent variable is log(contributions)

All explanatory variables are lagged by one period

The sample is restricted to charities with a 3-star rating in 2015

Ratings are fixed at the rating a charity received in 2016, and used as an instrument for the actual ratings they receive

Small charities are those with net assets less than $1 million (10th percentile)

Large charities are those with net assets greater than $57 million (90th percentile)

Standard errors (in parentheses) clustered at the charity level

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.7: Administrative Expense score - 10 point threshold

	Average bunching response	Observations
All Years	-0.204	58749
	(0.142)	
2019	-0.301	5750
	(0.273)	
2018	-0.599**	6728
	(0.291)	
2017	-0.245	6121
	(0.307)	
2015	-0.437	5935
	(0.369)	

Notes: Bootstrapped standard errors are shown in parentheses
Lower bound is selected as 13
The sample excludes community foundations, museums and food and humanitarian relief charities, which are rated on a different scale
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.8: Probability of Being Rated - Logit Model

Number of paid employees	0.0549
	(0.0686)
Number of volunteers	0.177***
	(0.0310)
Total contributions	0.170***
	(0.0596)
Fundraising expenses	0.129***
	(0.0405)
Total wage bill	-0.350***
	(0.0880)
Observations	27529
R-Squared	.0208

Notes: The dependent variable equals 1 if a charity is rated at some point in the sample period, and 0 if they remain unrated
All explanatory variables are in logs
Standard errors (in parentheses) clustered at the charity level
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

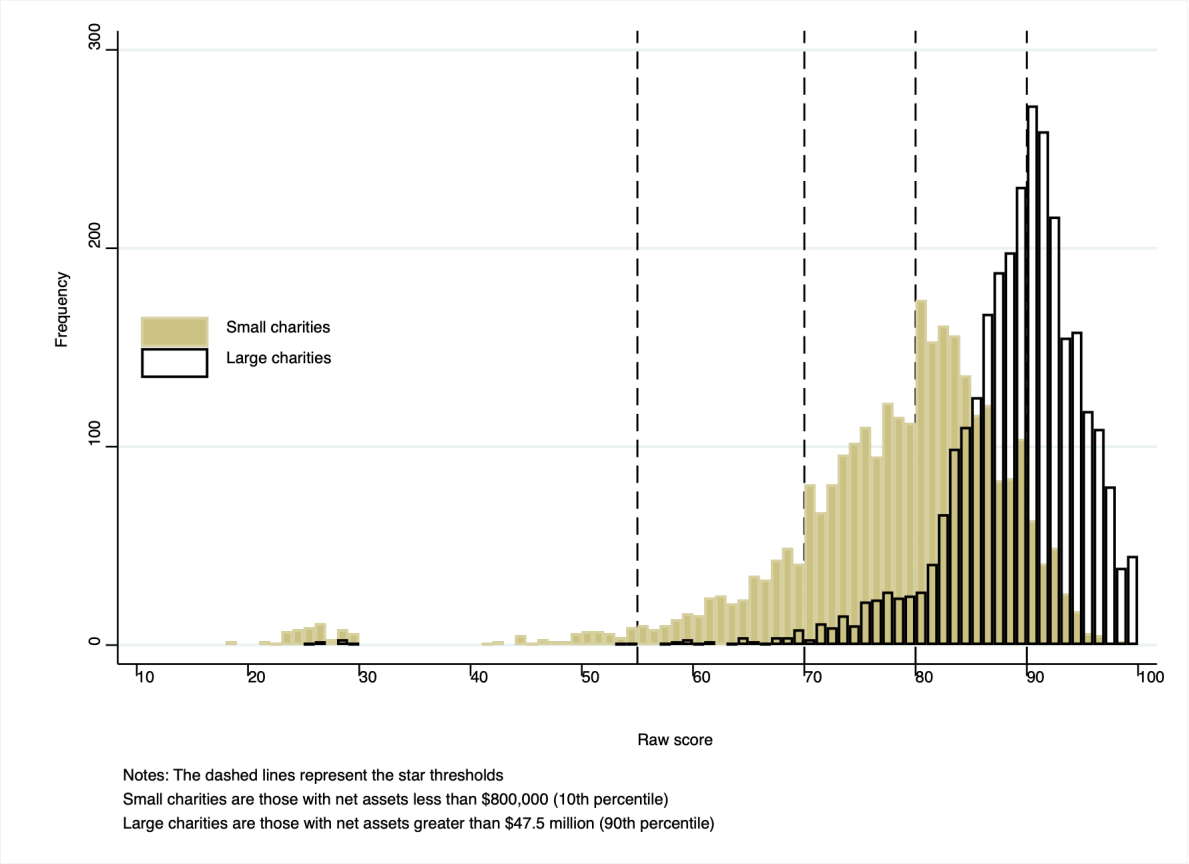
Table B.9: Rated for the First Time

	w_{te}*100	w_{pe}*100
Newly Rated * Period t	-0.158	-0.187
	(0.135)	(0.208)
Newly Rated * Period t+1	0.121	0.086
	(0.120)	(0.192)
Newly Rated * Period t+2	0.006	-0.125
	(0.133)	(0.215)
Charity Fixed Effects	Yes	Yes
Time Fixed Effects	Yes	Yes
Observations	43611	43611
R-Squared	.0002	.0002

Notes: Standard errors (in parentheses) clustered at the organization level
The omitted reference period is the period prior to first being rated
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C Appendix Figures

Figure C.1: Bunching in the Overall Score by Charity Size - 2017-19



Notes: The dashed lines represent the star thresholds
Small charities are those with net assets less than $800,000 (10th percentile)
Large charities are those with net assets greater than $47.5 million (90th percentile)

Figure C.2: Bunching in the Overall Score by Tax Preparer - 2017-19

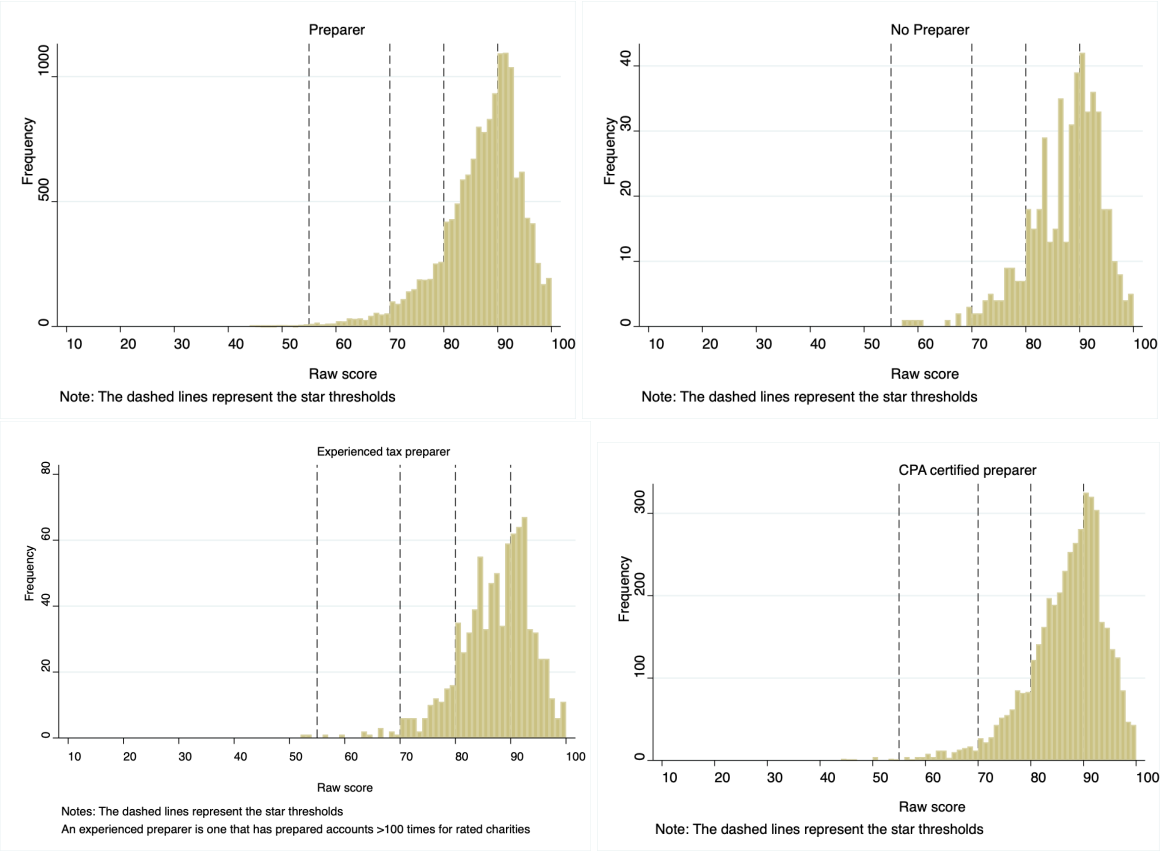


Figure C.3: Bunching in the Overall Score by Contribution Reliance

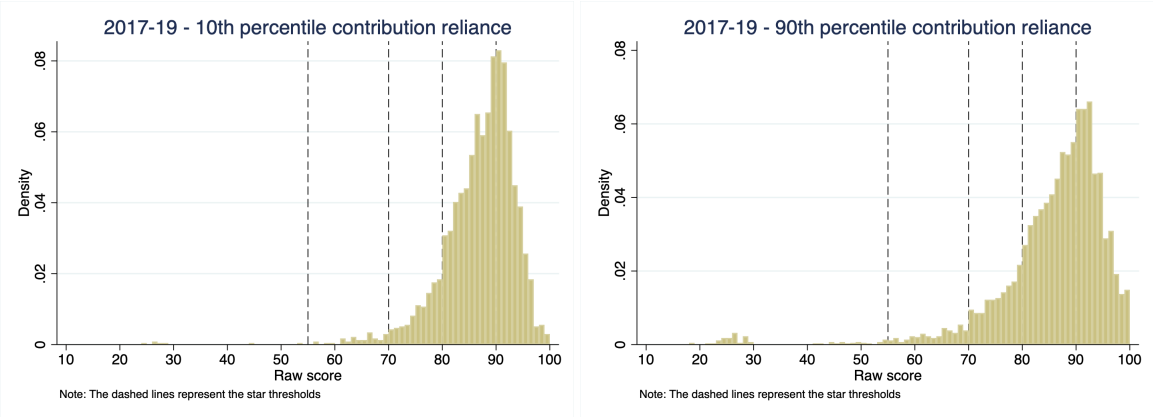


Figure C.4: Bunching in Administrative Expense Percentages - 2002-19
a) Food and Humanitarian Relief b) Housing and Human Services

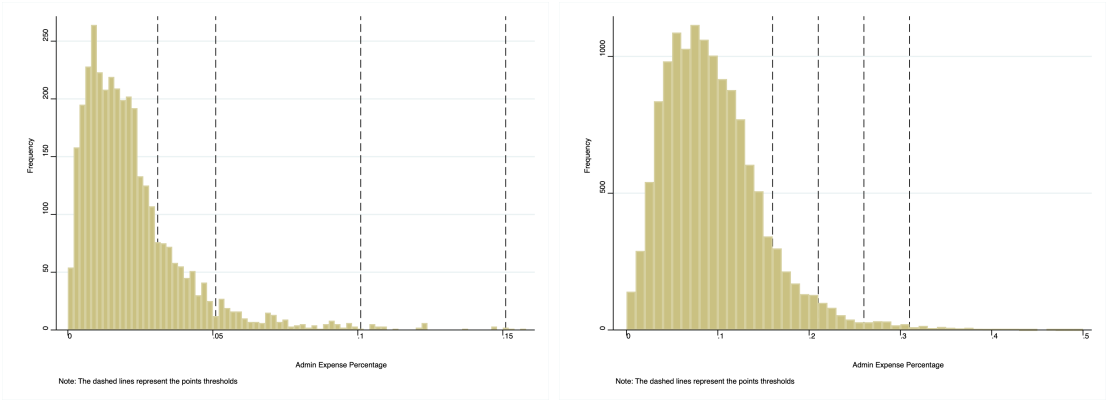


Figure C.5: Bunching from Above?

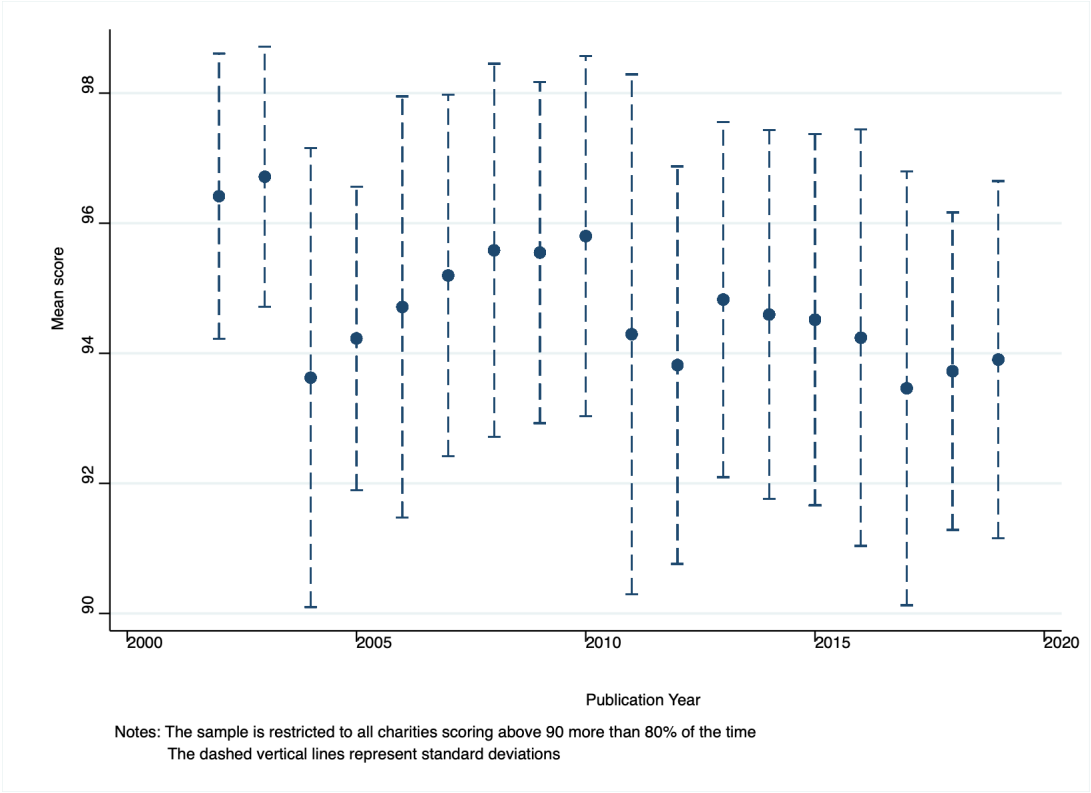


Figure C.6: Distribution of pe_{te} for Rated vs Unrated Charities

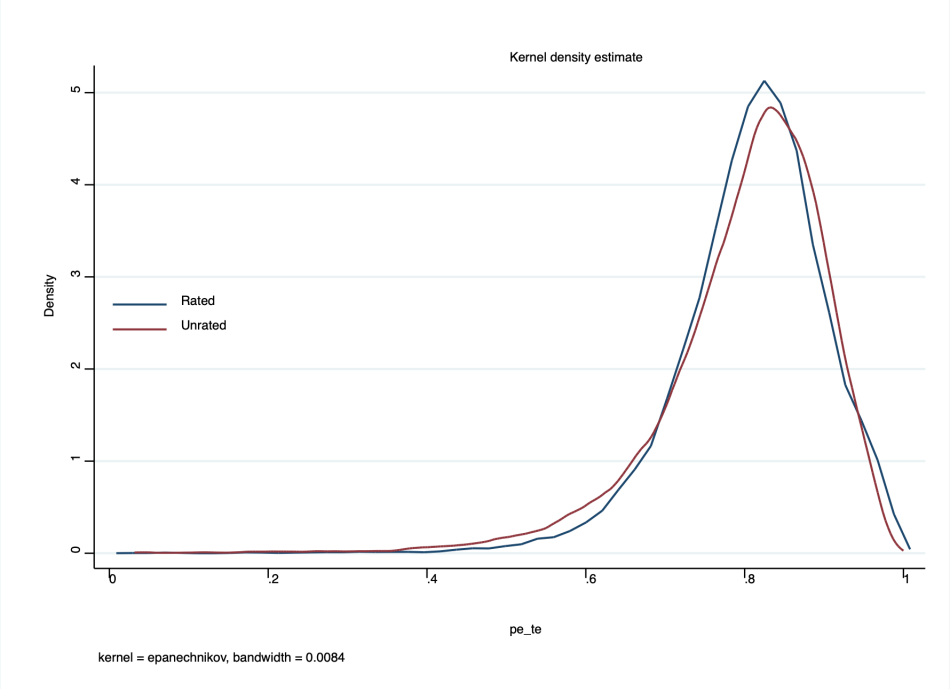


Figure C.7: Distributions of w_{te} and w_{pe} for Rated vs Unrated Charities

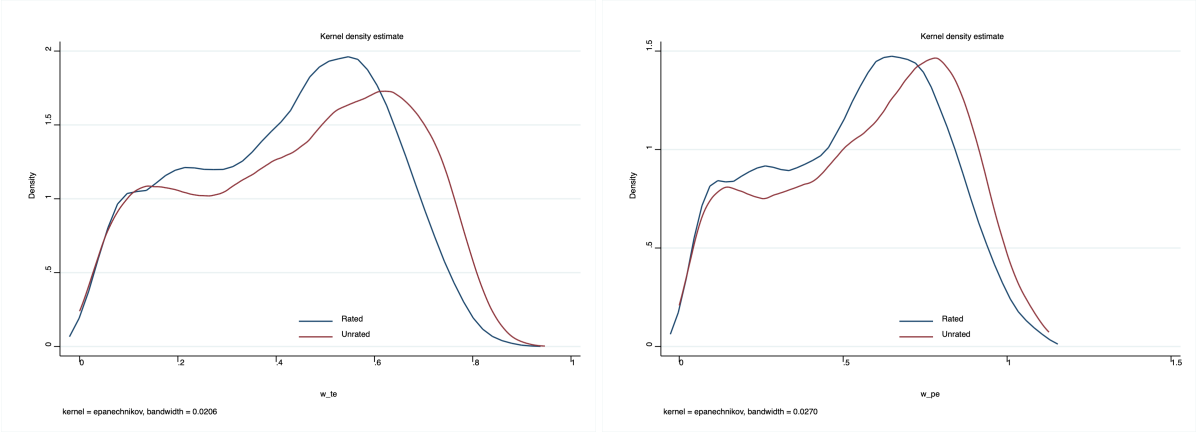


Figure C.8: Corrected Scores (2017-19) - Preparer

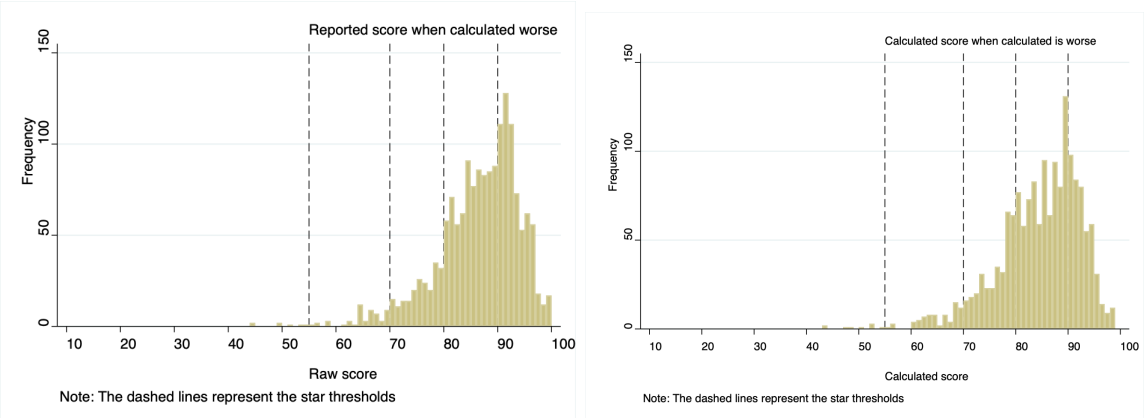


Figure C.9: Corrected Scores (2017-19) - No Preparer

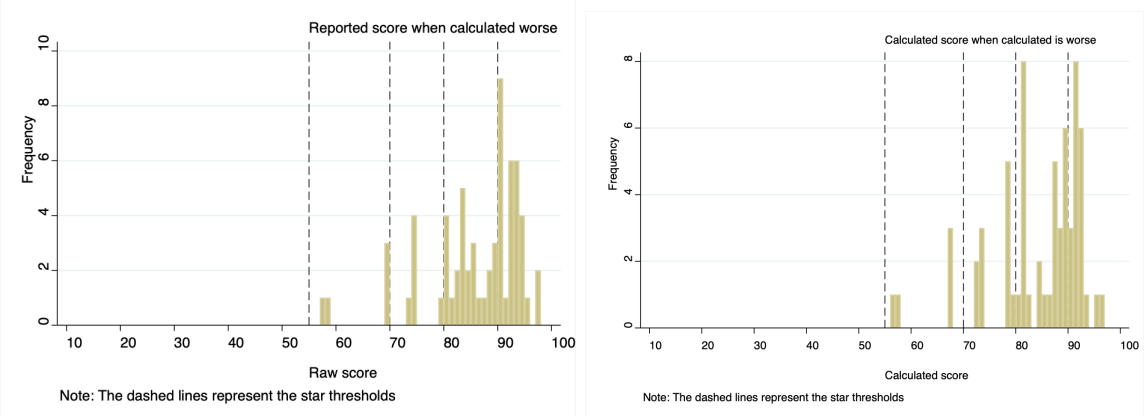
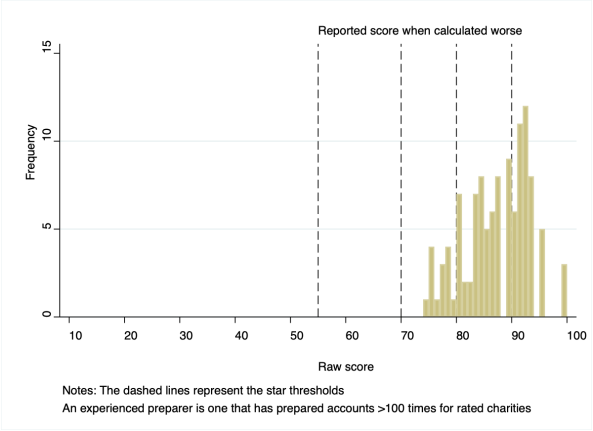


Figure C.10: Corrected Scores (2017-19) - Experienced Preparer



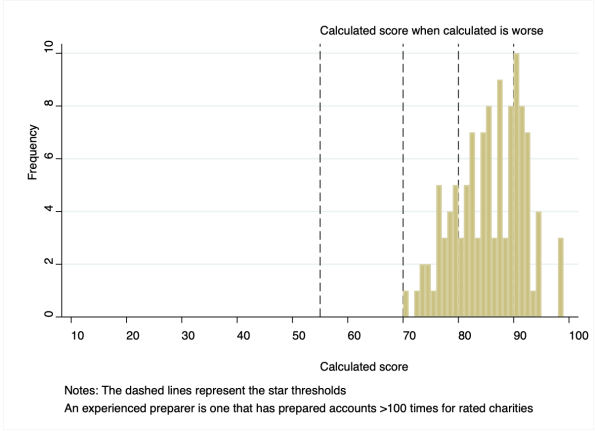


Figure C.11: Corrected Scores (2017-19) - CPA-Certified Preparer