

Georgia State University

ScholarWorks @ Georgia State University

AYSPS Dissertations

Andrew Young School of Policy Studies

Summer 8-11-2020

Information Asymmetry, Organizational Performance, and Private Giving: Can Performance Ratings Build Trust in Nonprofits?

Iurii Davydenko
ydavydenko1@gsu.edu

Follow this and additional works at: https://scholarworks.gsu.edu/ayspss_dissertations

Recommended Citation

Davydenko, Iurii, "Information Asymmetry, Organizational Performance, and Private Giving: Can Performance Ratings Build Trust in Nonprofits?." Dissertation, Georgia State University, 2020.
https://scholarworks.gsu.edu/ayspss_dissertations/4

This Dissertation is brought to you for free and open access by the Andrew Young School of Policy Studies at ScholarWorks @ Georgia State University. It has been accepted for inclusion in AYSPPS Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

ABSTRACT

INFORMATION ASYMMETRY, ORGANIZATIONAL PERFORMANCE, AND PRIVATE GIVING: CAN PERFORMANCE RATINGS BUILD TRUST IN NONPROFITS?

By

IURI DAVYDENKO

August 2020

Committee Chair: Dr. Dennis R. Young

Major Department: Public Management and Policy

Nonprofit performance report cards, such as charity ratings, have evolved in the third sector as an attractive tool for addressing accountability concerns and improving the sector's effectiveness and efficiency. These performance monitoring services intend to increase the quality of philanthropy by helping donors allocate contributions to high-quality charities and getting organizations to improve their performance. However, we know little about how performance report cards as a policy instrument fulfill their expectations in the nonprofit sector.

This research offers a comprehensive study of charity ratings that addresses three sets of questions. First, it explores the information content of charity ratings and assesses the degree of coherence among performance grades assigned by different rating services. The analysis of data shows that the informational content of charity performance ratings is lower than it appears on face value, and competing rating systems often send mixed signals to donors.

Second, it examines whether and how conventional metrics embedded in charity ratings, particularly composite ratings and overhead spending ratios, influence perceived performance, trust, and giving decisions in individual donors. The findings show that individuals consider both ratings and overhead ratios when making decisions but give the ratings more weight. The study also reveals distinct patterns in donor reactions to low and high values on each of the two measures, interactions between them, and a moderating role of altruism, general trust, and mission valence.

Finally, the study investigates how rated nonprofits respond to their ratings. It proposes that a public charity will react to an exogenous shock - the release of its charity rating by improving its measured performance, especially if it (1) initially gets a poor rating, (2) is in a highly competitive subfield, (3) relies more heavily on donations. The empirical tests show that public charities only respond in a limited way to being publicly rated, meaning limited effectiveness of the existing tool to elicit performance improvements in nonprofits. At the same time, the statistically and practically significant findings for the charities that initially receive the lowest ratings show that third-party nonprofit performance monitoring has some potential.

INFORMATION ASYMMETRY, ORGANIZATIONAL PERFORMANCE, AND PRIVATE
GIVING: CAN PERFORMANCE RATINGS BUILD TRUST IN NONPROFITS?

BY

IURI DAVYDENKO

A Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree

of

Doctor of Philosophy

in the

Andrew Young School of Policy Studies

of

Georgia State University

Georgia State University

August 2020

Copyright by
Iurii Davydenko
2020

ACCEPTANCE

This dissertation was prepared under the direction of the candidate's Dissertation Committee. It has been approved and accepted by all members of that committee, and it has been accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Public Policy in the Andrew Young School of Policy Studies of Georgia State University.

Dissertation Chair:	Dr. Dennis R. Young
Committee:	Dr. Gregory B. Lewis Dr. Theodore H. Poister Dr. Michael K. Price Dr. Gregory D. Streib

Electronic Version Approved:
Sally Wallace, Dean
Andrew Young School of Policy Studies
Georgia State University
August 2020

DEDICATION

To those who never stop striving to be better themselves.

ACKNOWLEDGEMENTS

I want to express my sincere gratitude and appreciation to the following people who contributed immensely to the successful completion of my doctoral studies and dissertation.

I am incredibly grateful to my stellar dissertation committee – Dennis R. Young, Gregory B. Lewis, Theodore H. Poister, Michael K. Price, and Gregory D. Streib – not only for accepting the mission of serving on the committee but for your time, patience, enthusiasm, academic brilliance, and invaluable advice.

I would also like to extend my deepest gratitude to our Ph.D. Program Director Christine Roch, for believing in my determination to see this project to its successful completion and for her support.

Special thanks to Matt Arp, Elsa Gebremedhin, and Amber Slyter for all the help I received from you on the administrative side of the process. This accomplishment would not have been possible without your hard work and assistance.

TABLE OF CONTENTS

	Page
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: CONTENT AND COMPARABILITY OF CHARITY RATINGS	9
2.1 Introduction	9
2.2 Overview of Nonprofit Rating Agencies	12
2.3 Content and Comparability of Charity Ratings: Hypotheses	25
2.4 Data and Methodology	28
2.5 Findings	30
2.6 Conclusions and Implications	37
CHAPTER 3: INDIVIDUAL DONOR RESPONSE TO CHARITY RATINGS	39
3.1 Introduction	39
3.2 Performance, Trust, & Individual Donor Behavior	42
3.3 Performance Ratings and Donor Reactions: Research Hypotheses	46
3.4 Methodology	54
3.5 Measurement	57
3.6 Findings	59
3.6.1 Perceptions of Overall Performance	59
3.6.2 Perceived Impact	61
3.6.3 Perceived Efficiency	64
3.6.4 Donor Trust	67
3.6.5 Donation Preference	68
3.7 Conclusions and Policy Implications	73
3.8 Limitations and Future Research Directions	75
CHAPTER 4: PUBLIC CHARITY RESPONSE TO PERFORMANCE RATINGS	77
4.1. Introduction	77
4.2 Literature Review	80
4.2.1 Do Organizations Change Behavior in Response to Ratings? Evidence from Education, Healthcare, and the Corporate Sector	80
4.2.2 Theory of Organizational Response to External Performance Monitoring	83
4.2.3 Public Charity Response to Third-Party Performance Ratings: Theory and Hypotheses	92
4.3 Data	97

4.4 Methodology	100
4.5 Findings	105
4.6 Conclusion.....	117
4.7 Limitations and Future Research Directions	118
APPENDIX A.....	121
APPENDIX B	124
APPENDIX C	129
BIBLIOGRAPHY.....	133
VITA	143
DISCLOSURE STATEMENT	144

LIST OF TABLES

Table 2. 1: Transparency seals from GuideStar.....	13
Table 2. 2: Summary of defining characteristics of nonprofit rating agencies.....	14
Table 2. 3: Charity Navigator's overall rating and overall score	21
Table 2. 4: Qualitative interpretation of the CN's star ratings	21
Table 2. 5: Relative contributions of the Financial Score and Accountability and Transparency Score in explaining the variation in the Overall Score	32
Table 2. 6: Hierarchical linear regression using R ² -based forward model selection.....	33
Table 2. 7: Relative contributions of the reported Program Expense Ratio and Fundraising Efficiency in explaining the variation of the Charity Watch grades.....	33
Table 2. 8: Pearson correlation coefficients between the CN and CW performance grades.....	35
Table 2. 9: Cross-rater consistency in rating grades	36
Table 3. 1: Experimental conditions	56
Table 3. 2: Differences in perceived overall performance across treatments	60
Table 3. 3: Within-group differences.....	61
Table 3. 4: Perceived impact across treatments (ordinal logistic regression).....	62
Table 3. 5: Perceived efficiency across treatments (ordinal logistic regression).....	65
Table 3. 6: Donor trust across treatment groups	68
Table 3. 7: Willingness to donate across treatments.....	70
Table 3. 8: Willingness to donate across treatments (Low-Rating-Low-Overhead)	70
Table 3. 9: Willingness to donate across treatments (High-Rating-High-Overhead).....	71
Table 3. 10: Willingness to donate across treatments.....	72
Table 4. 1: Charity Navigator's grading scales.....	99
Table 4. 2: Coding scheme for time periods in the ratings dataset.....	101
Table 4. 3: Agency measured performance after initial Charity Navigator's rating	106
Table 4. 4: Agency measured performance after initial Charity Navigator's rating depending on the value of the initial rating (reference group for Initial Rating = "Good").....	109
Table 4. 5: Agency measured performance after initial Charity Navigator's rating for initially top-rated charities.....	111
Table 4. 6: Computed average competition by category of charitable activity (sorted from the highest competition to the lowest)	112
Table 4. 7: Moderating effect of the competition across fields of activity	113
Table 4. 8: Moderating effect of the share of public contributions on agency response to being rated.....	116
Table A. 1: Grade Conversion Scheme.....	123
Table B. 1: Charities selected for the experiment and performance report cards presented	124
Table B. 2: Demographic characteristics of the study sample by treatment groups.....	126
Table B. 3: Question items for the measure of trust in a charity	126
Table B. 4: Question items for the measure of dispositional (personality) trust	126
Table B. 5: Question items for the measure of altruism	127
Table B. 6: Question items for the measure of trust in nonprofits in general.....	127

Table B. 7: Question items for the measure of perceived performance.....	128
Table B. 8: Question items for the measure of emotional utility.....	128
Table B. 9: Question items for the measure of familial utility	128
Table C. 1: Charity program expenses for 8640 charities rated during 2000-2018 (percent) ...	129
Table C. 2: Charity program expenses average annual growth for 8640 charities rated during 2000-2018 (percent).....	130
Table C. 3: Distribution of initial overall ratings for 8640 charities rated during 2000-2018 (measured in stars)	131
Table C. 4: Descriptive statistics for the distribution of the mean share of public contributions in nonprofit agency income portfolios for 8640 charities rated during 2000-2018.....	132

LIST OF FIGURES

Figure 2. 1: Summary of a report card from BBB’s give.org	17
Figure 2. 2: Summary of a report card from Charity Watch.....	19
Figure 2. 3: Summary of a report card from Charity Navigator.	20
Figure 2. 4: Summary of a report card from Impact Matters.....	24
Figure 2. 5: Inter-rater consistency in charity performance grades	36
Figure 3. 1: Perceptual Determinants of Charitable Giving	44
Figure 3. 2: Differences in perceived overall performance across treatments.....	60
Figure 3. 3: Probability differences in perceived impact across treatments	64
Figure 3. 4: Probability differences in perceived efficiency across treatments	67
Figure 3. 5: Donor trust across treatment groups.....	68
Figure 3. 6: Donor willingness to donate across treatment groups.....	69
Figure 4. 1: Research model	97
Figure A. 1: The distribution of the 102,534 charity ratings for the 8640 charities rated by the Charity Navigator.....	121
Figure A. 2: The distribution of the 595 charities rated by the Charity Watch	121
Figure A. 3: The distribution of performance grades assigned by Charity Navigator.....	122
Figure A. 4: The distribution of performance grades assigned by Charity Watch	122
Figure A. 5: Overall published rating scores against overall recalculated scores before and after cleaning the dataset.	123
Figure C. 1: Charity program expenses for 8640 charities rated during 2000-2018 (percent)..	129
Figure C. 2: Charity program expenses average annual growth for 8640 charities rated during 2000-2018 (percent).....	130
Figure C. 3: Distribution of initial overall ratings for 8640 charities rated during 2000-2018 (measured in stars)	131
Figure C. 4: Distribution of the mean share of public contributions in nonprofit agency income portfolios for 8640 charities rated during 2000-2018.	132

CHAPTER 1: INTRODUCTION

In the spirit of American individualism and freedom to pursue common purposes, for many decades, the voluntary sector in the U.S. has enjoyed unprecedented levels of independence from government interference along with generous economic privileges. Costing the Treasury tens of billions of dollars in lost revenue annually, such autonomy and support reflected a deep societal belief in the nonprofit sector's high purpose and a great degree of public faith in its self-regulation capacity and accountability (Kelly, 1998; Salamon, 2012). However, in the last several decades, big concerns over charitable organizations' abuse of public trust have grown, inviting an increased governmental regulation of the sector and challenging its favorable nonprofit tax treatment.

Numerous investigations uncovering fraud, tax-avoidance, self-dealing, excessive costs, commercial activities, accounting manipulations, excessive accumulation of tax-exempt wealth, distortion, incomplete information, and other practices that do not benefit society, have undermined nonprofit credibility (Kelly, 1998). The most common charitable organizations' abuses involve overvaluing donated products and allocating fundraising expenses to the program category to keep the reported overhead cost low, disguised profit distribution through high salaries and benefits, excessive endowments, unfair competition, and unrelated business activities. Kelly (1998) argued that "Self-regulation has not worked up to this point" (p. 184), and "the era of giving charitable organizations the benefit of the doubt was over" (p.220).

The ongoing shift of public and government attention towards the increased scrutiny in the nonprofit sector have created an environment where, besides a more onerous regulatory burden, nonprofits are increasingly expected to prove their public value. Kelly (1998) argued that now charities "must explain themselves, demonstrate their service is worth the cost, and defend

their essential character” (p.185). Concerns over nonprofit performance accountability have pushed the adoption of performance measurement and monitoring in the nonprofit sector organizations (Carnochan, Samples, Myers, & Austin, 2014). As a result, the recent decades have witnessed a proliferation of performance measurement and monitoring systems intended to equip nonprofits with tools that should demonstrate what difference the dollars entrusted to them make and their key stakeholders with means to make informed choices. One of these instruments – internet technology-empowered charity ratings as an implementation of performance monitoring report cards – is the object of interest in this research.

Performance accountability tools in the nonprofit sector are supposed to improve performance through two mechanisms. First, performance information should inform key nonprofit stakeholders, particularly funders, about organizational or program quality, so that they could make justified decisions regarding their willingness to financially or otherwise support organizations. This mechanism would contribute to better and more efficient outcomes by helping markets to “weed out” poorly performing operations (Gormley & Weimer, 1999). Second, nonprofits can (and, advocates of performance measurement assert, should) use information produced in the process of performance measurement to help manages and governing boards identify opportunities for improvement and survive competition for resources (Poister, Aristigueta, & Hall, 2014; Wholey & Hatry, 1992).

Despite the pressure towards greater performance accountability and improvement, the embracement of performance measurement in nonprofits has been easier said than done. The nonprofit performance revolves around an organization’s ability to convert its inputs efficiently to mission-related, social outcomes (Gormley & Weimer, 1999), while, at the same time, addressing the challenge of survival and sustainability. Measuring multidimensional nonprofit

performance is notoriously difficult, and organizations often see it as a resource drain and an unjustified burden (MacIndoe & Barman, 2012). As a result, organizations have been slow to adopt performance measurement and tend to implement it rather superficially. The general public's limited ability to obtain and interpret complex performance information across diverse entities and a variety of measurement approaches further diminish the promise of performance measurement to alleviate information asymmetry in the nonprofit sector and facilitate performance-based accountability (Herzlinger, 1995).

Organizational report cards are a policy instrument that could relieve nonprofits of the burden of performance measurement and information dissemination while holding the potential to strengthen bottom-up accountability and self-regulation. Gormley and Weimer (1999) define organizational report cards as “a regular effort by an organization to collect data on two or more other organizations, transform the data into information relevant to assessing performance, and transmit the information to some audience external to the organizations themselves” (p.3). The researchers highlight the following three elements that distinguish report cards from other instruments of performance accountability. First, they are external assessments and thus carry the cost of performance measurement. Second, they assume regular assessment. Third, they are designed for an external audience to facilitate easy access, comprehension, and comparison of performance across organizations and over time.

A common approach in report cards is to transform performance information in ratings or rankings. Many such ratings and rankings have recently spread in education, healthcare, childcare, restaurant, finance, and other industries to inform customer choices and rated organizations' behavior (Gormley & Weimer, 1999; Jin & Leslie, 2003; Lewis, 2014; Sharkey & Bromley, 2014). By taking advantage of the advancements in information technologies,

charity/nonprofit ratings have evolved in the last two decades. These report cards focusing on nonprofit organizations' performance accountability are produced by third-party evaluators – charity/nonprofit rating agencies also commonly addressed as charity watchdog organizations. Like credit rating agencies that evaluate potential borrowers in the private and public sectors (such as Moody's, Fitch, or Standard & Poors), charity raters are independent, private, and self-sustaining organizations. Numerous watchdog agencies grade and distribute information on thousands of nonprofits in the US and abroad (Rowe, 2012) in order to facilitate nonprofit accountability. Some of them produce and disseminate nonprofit performance report cards in accordance with the definition of this policy tool¹.

Charity ratings are a potentially powerful monitoring instrument for facilitating nonprofit performance accountability and stimulating organizational improvements. Charity raters collect and analyze information about NPOs' performance using objective criteria and deliver it to the public in a convenient for decision-making letter or star grade scale. Unlike commercial service agencies, however, charity raters are often nonprofit organizations themselves funded through voluntary contributions. The nonprofit status might indicate not only the public's demand for such performance information but also the potential ability of the nonprofit sector to produce at least partial remedies for its "voluntary failure²." In most cases, access to charity ratings and the underlying data is free or relatively low cost. With such an approach to the design and funding of the nonprofit report cards, charity raters effectively make their performance assessments accessible to individual donors.

¹ Not all charity ratings are performance report cards as defined by (Gormley & Weimer, 1999), but in this research, the terms "nonprofit report cards" and "charity ratings" will be used interchangeably.

² Nonprofit inability to address market failures due to covert distributions, productive inefficiencies, and other reasons

Performance measurement experts have often criticized charity ratings. These criticisms are justified, and charity performance report cards may not have the potential to address all nonprofit accountability needs or may fail to deliver according to expectations. Gormley and Weimer (1999) argued that “the design and use of organizational report cards involve a number of generic problems that undercut their value as a policy instrument” (p.7). They pointed to three potential challenges that may weaken report cards: assessment problems, consumer reception problems, and organizational response problems. Assessment problems reflect limitations of measurement, as failure to assess performance comprehensively may affect usefulness of the information report cards provide to their users and lead to undesirable behaviors. Reception and organizational response problems are related to a variety of potentially dysfunctional responses of consumers of the information and targeted organizations. The researchers explained that reception problems may arise due to “weak motivation”, “cognitive limits”, and “informational inequalities” (p. 15), while response problems may include “nonparticipation”, “cream skimming”, “manipulating the numbers”, and “blaming the messenger” (p. 13).

Nonprofit performance report cards have been available to the public for more than three decades, and the field continues to evolve. Charity ratings can be a useful instrument for improving self-regulation of the nonprofit sector and strengthening its credibility. However, the research on nonprofit ratings is fragmented and inconclusive. It offers limited insights into how performance report cards as a policy instrument fulfill its expectations in the nonprofit sector, particularly in terms of its impacts on the behavior of intended audiences. Given the increasing coverage and salience of ratings in the nonprofit environment and the limited research available on this topic, many essential questions merit theoretical and empirical examination. By focusing on some of these questions, this research is organized into three parts:

1. *Informational Content and Comparability of Charity Ratings*. This descriptive section determines the main drivers of the variation in charity ratings and assesses the degree of coherence among assessments delivered by different charity rating agencies.
2. *Individual Donor Response to Third-Party Charity Ratings*. Using experimental data, I examine whether and to what extent donors respond to information about charity ratings.
3. *Public Charity Response to External Performance Ratings*. Using observational data, I examine how rated nonprofits respond to their ratings.

A few additional caveats are in order. Because the nonprofit sector is represented by different categories of organizations with diverse purposes, underlying business models, legal structures, and management practices (Salamon, 2012), it is important to clarify the boundaries of the population of nonprofits that this research is relevant to. Following Hansmann's and Weisbrod's classical ideas about the rationale for nonprofit organizations, the interest in this research is on voluntary donative organizations that produce services with public good characteristics and those whose operations can be characterized by information asymmetry between nonprofits and their donors. From the legal perspective, these are voluntary corporations that are bound by the nondistribution constraint, generate public benefits, and receive public support. In relation to the first criterion, this research applies to organizations that are legally prohibited from distributing surplus among their founders. The second criterion further restricts the population of interest to *public-serving (charitable) organizations* that "benefit an indefinite class of individuals" (Powell & Steinberg, 2006, p. 2) as opposed to *mutual benefit* or, in Salamon's terms, *member-serving organizations* (Salamon, 2012). The third criterion requires

that a nonprofit has substantial support from individuals, the importance of which follows from the theory of nonprofit demand and, on the other hand, from a greater severity of information asymmetry that individuals suffer in comparison to institutional stakeholders.

One setting where this research would be especially relevant is online giving. Even though online giving only accounts for about ten percent of the total giving (E. Brown & Martin, 2011; MacLaughlin, 2015; NPTrust, 2015), it has a strong potential for growth. Today's broad spread of the internet and mobile applications, development of e-payment options, deep penetration of online social networks into individuals' personal and professional lives, the rise of big data analytics, and development of highly sophisticated algorithms of behavioral analysis online allow business managers to achieve impressive business goals³. Such internet tools could greatly facilitate online fundraising in the nonprofit business and help public charities attract substantial amounts by reaching large online audiences and providing them with the right information. Research, in turn, shows that online giving has grown persistently during the last decade at substantially higher rates than giving through traditional channels (NPTrust, 2015; Rovner, Loeb, McCarthy, & Johnston, 2013). Virtually every charity now has a website, so, from the policy perspective, charity ratings might be an efficient and less intrusive, incentive-based regulation tool (Gormley & Weimer, 1999). Relatively little scholarship providing insights into determinants of online giving behavior has been produced. The question of how the digital medium can complement physical interaction between organizations and individuals that care about and willing to support a cause to reduce uncertainty and facilitate trust is of crucial

³ What Can Companies Predict From Your Digital Trail? <http://www.npr.org/2015/09/14/440305167/what-can-companies-predict-from-your-digital-trail>

importance. Charity ratings are inherently an online tool, which might be part of a productive online relationship between nonprofits and their funders.

The nonprofit sector is responsible for a significant share of the whole economy and many socially essential functions. Public perceptions of nonprofit organizations' goals and performance may have substantial implications for future confidence in the third sector entities, their role, resource base, structure, and viability. Understanding the behavior of nonprofit organizations and their constituents conceptually and empirically might suggest which tools can effectively push the sector to perform at the maximum of its potential. This research is looking to gain insights on important aspects of donor and organization behavior in the presence of charity performance rating information.

CHAPTER 2: CONTENT AND COMPARABILITY OF CHARITY RATINGS

2.1 Introduction

Organizational report cards have substantially proliferated in many markets and areas of public life over the past few decades. Scholarly research has documented the growth of performance report cards and their impact on school choice, funding, and expenditures (Figlio & Kenny, 2009; Jin & Whalley, 2007a, 2007b); hospital choice, revenue, and patient volume (Pope, 2009), and a variety of other outcomes in the private and public sectors (Gormley, 2003; Gormley & Weimer, 1999; Johnson & Kriz, 2002; Zhe Jin, Kato, & List, 2010). Charity ratings have also noticeably proliferated, and evidence is mounting that they may influence the behavior of various nonprofit stakeholders too (A. L. Brown, Meer, & Williams, 2014; Chhaochharia & Ghosh, 2008; Gordon, Knock, & Neely, 2009; Sloan, 2009).

Despite the strong theoretical rationale for report cards in facilitating the bottom-up accountability and their widespread proliferation, this policy instrument's designs are not always effective (Gormley & Weimer, 1999). One of the major reasons performance report cards may not live up to the expectations originates in methodological challenges of performance evaluation they face. Gormley and Weimer (1999) explained, "Fundamentally, these problems arise because of limitations in data and theory: not all relevant variables can be measured, and theoretical links between variables that can be measured and those that are conceptually appropriate are often weak" (p. 7). In the nonprofit literature and media space, this issue is framed as a matter of "watching the watchdogs" (Eng, 2011; Kelly, 1998) or "rating the raters" (National Council of Nonprofit Associations and the National Human Services Assembly, 2005).

The notion of performance measurement in the nonprofit sector revolves around multiple dimensions of organizational and program performance, such as effectiveness, efficiency, quality, equity, etc. (Poister et al., 2014). Designing performance measurement systems, therefore, involves considering multiple categories of measures in the chain between inputs and outcomes. Among others, these include measures of inputs, outputs, productivity, efficiency, service quality, customer satisfaction, outcomes/impacts, cost-effectiveness. Because comprehensive measurement is unfeasible, system designers face the challenge of choosing which measures to use and which to ignore (Gormley & Weimer, 1999; Poister et al., 2014). Gormley and Weimer (1999) warned that in designing performance report cards “the choice is often between an imperfect report card and no report card at all” (p. 10).

Since performance report cards aim to reduce multidimensional organizational performance to a user-friendly, comprehensible measure or set of measures (e.g. overall performance scores or ratings), the potential of a report card to alleviate information asymmetry, its credibility, and effectiveness, therefore, depend on the amount of information incorporated in its composite performance indicators. However, it is not always clear what pieces of information such systems truly reflect. For instance, research from health-care industry shows that even when a rater claims it uses multiple factors in its evaluation methodology, the ranking can be driven almost entirely by a single measure (Pope, 2009). Do nonprofit raters fall in the same trap? Because charity ratings have also been criticized for reliance on overly simplistic measures of financial efficiency pulled from 990 forms, particularly the overhead ratio (Lowell, Trelstad, & Meehan, 2005; National Council of Nonprofit Associations and the National Human Services Assembly, 2005), the first question this study seeks to answer is: what drives variation in charity

ratings? By exploring this question, this research will investigate whether charity ratings incorporate more performance information than using just efficiency ratios would reveal.

Another issue with the report cards in the category of assessment problems relates to consistency among various raters and choosing among them. Different raters that emerged at different times use different rating approaches. Various raters often claim they do a better job evaluating charities. Not only nonprofit agencies, their associations, and the media sometime attack charity ratings (Kelly, 1998; Lowell et al., 2005), charity watchdogs criticize each other (Charity Watch, 2012; O'Donnell, 2012). USA Today, for instance, questioned the credibility of Better Business Bureau's ratings for approving of charities that get an F from The American Institute of Philanthropy's Charity Watch and a zero-star rating from Charity Navigator (O'Donnell, 2012). The Charity Watch, on its website, brings users' attention to the fact that F rated charities get certified by Independent Charities of America (Charity Watch, 2012) and publishes scathing criticism of Great Nonprofits – a community-sourced rater (GreatNonprofits, 2015).

Existence of different charity certifiers using different assessment methodologies is not a problem per se. In fact, competition among raters may have benefits (Lizzeri, 1999). At the same time, if the assessments of organizational quality are inconsistent across different rating service providers, users may face the question of choice among them, and the cost of information search as well as uncertainty may increase. As an example, a study of information intermediaries in municipal debt markets shows that receiving split ratings by municipal borrowers affects the cost of capital (Johnson & Kriz, 2002). In other word, this means that market participants consider multiple ratings in their decision making. An increased information search and processing cost with respect to nonprofit performance ratings may lower the usefulness and use of charity

watchdogs as quality certifying services for users of such information and decrease the policy instrument's regulatory potential. Therefore, the second question in this research asks: How consistent are the performance assessments delivered by competing charity raters to nonprofit constituents? The answers to these questions are not obvious and will inform our understanding of the informational role of nonprofit quality certifiers, including their rating methodologies, the content of ratings, and interrater consistency.

The next section provides a brief overview of the U.S. based nonprofit report cards. The following section formulates hypotheses about content and comparability of charity ratings. The Data and Methodology describes the empirical approach chosen for this analysis. The final section presents and discusses the findings, conclusion, and directions for further research.

2.2 Overview of Nonprofit Rating Agencies

This section provides brief descriptive profiles of the third-party assessment services in the nonprofit sector that are consistent with the definition of performance report cards. The literature and internet search aimed at identifying nonprofit information intermediaries yielded eight U.S. based institutions that supply nonprofit performance information to facilitate the third sector's accountability and improve decision making. This overview will focus on four of those evaluators, including (1) BBB Wise Giving Alliance, (2) Charity Watch, (3), Charity Navigator, and (4) Impact Matters. The other four identified watchdog organizations – GuideStar, Great Nonprofits, Give Well, and Forbes – are not included in the overview, as their information products are missing key data transformation elements that are inherent to performance report cards. Specifically, GuideStar and Give Well do not transform the information they collect in a form that facilitates easy interpretation and comparison of organizational performance. The Great

Nonprofits merely provides community-sourced qualitative reviews and reports aggregated community ratings as opposed to measuring objective performance. Finally, Forbes only ranks top 200 charities based on donations received along with reporting their total revenue, fundraising efficiency, and program spending.

Although GuideStar is not a watchdog and doesn't evaluate or rate charities, it's important to note it is the largest so far database of nonprofit data maintaining online profiles on 1.8 million IRS-recognized nonprofits and providing free access to Forms 990. Nonprofits can optionally provide additional information to their profiles on GuideStar and, based on the amount of provided information, earn one of its transparency seals – bronze, silver, gold, or platinum (GuideStar, 2020; GuidStar, 2020) as the Table 2.1 below shows:

Table 2. 1: Transparency seals from GuideStar




<p style="text-align: center;">Bronze</p> 	<p style="text-align: center;">Silver</p> 	<p style="text-align: center;">Gold</p> 	<p style="text-align: center;">Platinum</p> 
<p>Provide basic information about an organization to be found</p>	<p>Be transparent about Its finances to build trust</p>	<p>Share its goals and strategies</p>	<p>Share its quantitative measures of progress and results to show the difference it makes</p>

Table 2.2 below summarizes the key characteristics of the four rating agencies. Besides the general credentials of each agency, the table shows how their services fit the definition of the organizational report card and the differences in their measures, scale, scope, and cost of access for the public.

Table 2. 2: Summary of defining characteristics of nonprofit rating agencies

	(1) BBB Wise Giving Alliance	(2) Charity Watch (CW)	(3) Charity Navigator (CN)	(4) Impact Matters (IM)
Web Address	give.org	charitywatch.org	charitynavigator.org	impactmatters.org
Nonprofit Form	501(c)(3)	501(c)(3)	501(c)(3)	501(c)(3)
Scope	National	National	National	National
Total Revenue / Expenses	2,154,985/2,410,923 (FY 2018)	550,990/574,040 (FY 2018)	3,915,429/3,521,729 (FY 2018)	779,004/870,852 (FY 2018)
Established	1918/1977-2001	1992	2001	2017
Commitment to transparency on GuidStar.org	NA	Bronze	Platinum	Gold
Organizational focus	Yes	Yes	Yes	Yes
External assessment	Yes	Yes	Yes	Yes
External audience	Yes	Yes	Yes	Yes
Regularity	Yes	Yes	Yes	Yes
Data transformation	Yes	Yes	Yes	Yes
Reported Measure	Binary (Meets Standards / Standard is Not met)	Ten-point ordinal scale (letter grades from A+ to F)	Five-point ordinal scale (zero to four-star rating)	Five-point ordinal scale (one to five-star rating)
Number of rated organizations	1300	670	9000	1080
Type of rated organizations	Nationally soliciting charities	501(c)3, 501(c)4, 501(c)19, public support > \$1M	U.S. based registered 501(c)3, filing ≥ 7 years, revenue > 1M, public support ≥ \$0.5M, 40% revenue, fundraising exp. > 1%	Nonprofits that directly deliver services to people
Exceptions	Hospitals, houses of worship, and educational institutions	Houses of worship, PACs, clubs, colleges, hospitals, or other local institutions	Land trusts, hospitals & their foundations; universities; schools & their foundations; sorority & fraternity foundations; donor advised funds; charities with CN advisories, fiscal sponsors	Advocacy or systems change programs, religious organizations, community associations
Selection	Based on inquiries from constituents & charities	Does not accept requests from charities	If the criteria met, requests from the public are accepted through online voting	Consider requests from the public, but cannot rate individual nonprofits
Cost of Access to Report Cards for Users	Free of charge	Top Rated Charities – Free; Full Access - Paid Membership (\$50 annually)	Free of charge	Free of charge
Additional paid services	Accreditation Seal – Paid License	NA	NA	Impact Audits ⁴

⁴ According to the Forms 990 for FY2017 and FY2018, Impact Matters had programs service revenue for Impact Audits

BBB's Give.org

The first nonprofit watchdogs – the National Charities Information Bureau (NCIB) and the Philanthropic Advisory Service (PAS) of the Council of Better Business Bureau – date back to 1918 and 1977 correspondingly (Kelly, 1998). Each year, the two agencies audited 100-200 nationally soliciting charities (except hospitals, churches, and educational institutions) against their standards in finance, governance, transparency, and adherence to ethical and social norms. They distributed their findings on whether the audited charities were meeting the standards in the form of publications – educational brochures and summaries – as a free service to the public.

In 2001, the NCIB and PAS merged to form the BBB Wise Giving Alliance (BBB's Give.org) (Better Business Bureau, 2020) – a 501(c)(3) organization that seeks “to help donors make informed giving decisions by verifying if charities meet the 20 BBB Standards for Charity Accountability [and] strengthen charity practices” (BBB Wise Giving Alliance, 2019). Being the oldest nonprofit information intermediary, the BBB's Give.org assesses 1,300 nationally and over 10,000 locally soliciting charities. The evaluation reports are provided to the public at no cost through give.org (BBB Wise Giving Alliance, 2020). The charities that meet the BBB's standards can purchase a license to display a BBB Accredited Charity Seal in their fundraising materials.

BBB's Give.org evaluates charities against 20 standards in the five following groups:


1. Governance and oversight (five standards). This set of standards assumes that a charity has an active and independent volunteer board, as well as the institutions and procedures required to prevent it from potential self-dealing.


2. Measuring Effectiveness (two standards). This set of standards requires that an organization has measurable goals and a process of measuring its effectiveness in fulfilling its mission.
3. Finances (six standards). This is a set of standards for financial metrics, including program expenses ($\geq 65\%$), fundraising expenses ($\leq 35\%$), size of the unrestricted net assets available as well as budgeting and reporting practices.
4. Fundraising and Information Materials (five standards). Includes standards that ensure accuracy, completeness, and timeliness of a charity's communications to the public through solicitations, informational materials, annual reports, website disclosures, and marketing disclosures. It also includes the requirements related to donor privacy and addressing complaints.

When BBB's Give.org finds that a charity meets its 20 standards, it assigns the organization "Meets Standards" grade. Otherwise, the organizations receive the "Standards Not Met" or "Unable to Verify" status. An example of the summary of the report card is presented below:

CHARITY REVIEW Issued: December 2017 Expires: June 2020

American Humane

 **Accredited Charity**
Meets Standards



800-227-4645
1400 16th Street NW, Suite 360
Washington, DC 20036
<http://www.americanhumane.org>

[Full Report](#) [Share](#) [Print](#) [BBB Charity Standards](#)

Standards For Charity Accountability





















Governance	Finances	Fundraising & Info
1.  Board Oversight	8.  Program Expenses	15.  Truthful Materials
2.  Board Size	9.  Fundraising Expenses	16.  Annual Report
3.  Board Meetings	10.  Accumulating Funds	17.  Website Disclosures
4.  Board Compensation	11.  Audit Report	18.  Donor Privacy
5.  Conflict of Interest	12.  Detailed Expense Breakdown	19.  Cause Marketing Disclosures
	13.  Accurate Expense Reporting	20.  Complaints
Measuring Effectiveness	14.  Budget Plan	
6.  Effectiveness Policy		
7.  Effectiveness Report		

Figure 2. 1: Summary of a report card from BBB’s give.org. retrieved may 31, 2020 from <https://www.give.org/charity-reviews/national/animal-protection/american-humane-in-washington-dc-105>. Screenshot by author.

Overall, BBB’s Give.org relies on a comprehensive set of meaningful indicators of the nonprofit organization’s quality, but the binary scale of its reported summary measure that doesn’t allow one to distinguish between organizations within the passing or failing groups of charities appears to be its major shortcoming.

CharityWatch

The next oldest and, at the same time, smallest in terms of its evaluation capacity charity rater is the American Institute of Philanthropy (AIP) currently operating as CharityWatch. The agency rates mainly 501(c)3 public charities and some broadly soliciting 501(c)4 social welfare organizations and 501(c)19 veteran's organizations. Also, it focuses on nonprofits "that receive \$1 million or more of public support annually, are of interest to donors nationally, and have been in existence for at least three years" (CharityWatch, 2020b). The rating agency produces report cards for nearly 670 nonprofits and publishes them on its website. The service, however, is not entirely free of charge. The watchdog provides free public access to the report cards of its nearly 250 top-rated charities. Full access to CharityWatch reports requires purchasing a \$50 annual membership.

According to CharityWatch, its mission is to "maximize the effectiveness of every dollar contributed to charity by providing donors with the information they need to make more informed giving decisions" (Charity Watch, 2020). The report card published by the watchdog, however, focuses on organizational efficiency rather than effectiveness. It includes a CharityWatch Grade, financial measures determining the grade, information on whether the charity meets the CW's transparency and governance benchmarks, and additional descriptive information that may be material to donor decision making. The efficiency grade CharityWatch assigns each nonprofit it evaluates is a letter grade on an 11-point ordinal scale from A+ to F. The rater calculates the letter grade based on two financial efficiency measures – the program spending percentage and the cost to raise \$100. According to the CharityWatch's rating methodology, the agency assigns the final letter grade to a charity based on the average of the two measures using the CW's own scale after applying a system of adjustments that result from

in-depth evaluations of an organization’s financial reports, audited financial statements, tax forms, and other financial and nonfinancial documents and inquiries. The rating summary of the report card is shown below:

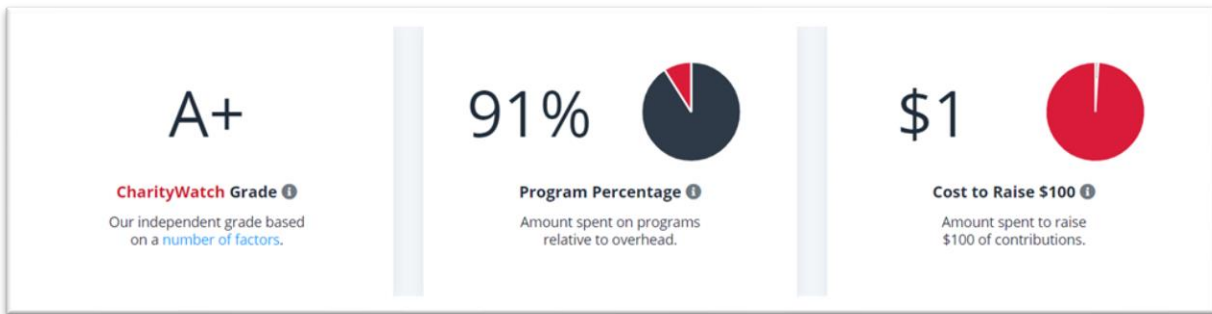


Figure 2. 2: Summary of a report card from Charity Watch. Retrieved May 31, 2020 from <https://www.charitywatch.org/charities/animal-welfare-institute>. Screenshot by author.

The adjustments that CharityWatch makes to the measures of efficiency and the resulting efficiency grades appear to be the hallmark of CharityWatch rating methodology. The evaluator makes adjustments so that they reflect the charity’s practices of treatment in-kind (non-cash) donations of goods and services (to capture charities’ potential inflation of the value of such assets), joint cost solicitation expenses (to capture a possible reporting of solicitation spending as program spending), and reserved assets (to capture excessive asset hoarding). CharityWatch claims that the extraordinary level of scrutiny they apply to their evaluations make their ratings “the most stringent in the sector” as opposed to “other charity information services [that] use simplistic or automated systems to generate ratings” (CharityWatch, 2020a). Nonetheless, CharityWatch clearly states their ratings reflect their opinion (CharityWatch, 2020c), whereas Lowell et al. (2005) criticized the CW ratings for lack of transparency and “gotcha” mentality.

Charity Navigator

Charity Navigator’s mission is to “make impactful philanthropy easier for all” by helping donors make informed giving decisions. It was established in 2001 and is currently the largest in terms of revenue and rating capacity nonprofit evaluator. It assigns performance ratings to approximately 9,000 registered as 501(c)(3) public charities with the annual revenue over \$1 million (including at least \$0.5 million in public support accounting for at least 40% revenue) that have been filing with the IRS for at least seven years. The criteria also require at least 1% of expenses to be allocated to fundraising. The agency provides its information services to the public free of charge and does not accept contributions from the rated charities (Charity Navigator, 2020c).

According to its methodology (Charity Navigator, 2020a), Charity Navigator assigns charities three numeric scores on a scale of 1-100 and three star-ratings on a five-point scale:

- Overall score and rating
- Financial score and rating
- Accountability and transparency score and rating.

All the mentioned scores and ratings are included in the main section of the CN’s report card as shown below:

	Score (out of 100)	Rating
Overall Score & Rating	73.90	★★☆☆
Financial	63.10	★☆☆
Accountability & Transparency	100.00	★★★★

Figure 2. 3: Summary of a report card from Charity Navigator. Retrieved May 31, 2020 from <https://www.charitynavigator.org/index.cfm?bay=search.summary&orgid=6082>. Screenshot by author.

The CN's overall score is obtained by applying the following mathematical transformation the two component scores:

$$Overall\ Score = 100 - \sqrt{\frac{(100 - Financial\ Score)^2 + (100 - Accountability\ \&\ Transparency\ Score)^2}{2}}$$

Each of the star-ratings is determined based on the corresponding performance score using the conversion scheme presented in Table 2.3:

Table 2. 3: Charity Navigator's overall rating and overall score

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

The CN's star ratings also have a qualitative interpretation as presented in Table 2.4.

Table 2. 4: Qualitative interpretation of the CN's star ratings

No. of Stars	Qualitative Rating	Description
★★★★★	Exceptional	Exceeds industry standards and outperforms most charities in its Cause.
★★★★☆	Good	Exceeds or meets industry standards and performs as well as or better than most charities in its Cause.
★★★☆☆	Needs Improvement	Meets or nearly meets industry standards but underperforms most charities in its Cause.
★★☆☆☆	Poor	Fails to meet industry standards and performs well below most charities in its Cause.
0-Stars	Exceptionally Poor	Performs far below industry standards and below nearly all charities in its Cause.
CN Advisory	No Rating	Serious concerns have been raised about this charity which prevents the issuance of a star rating

Charity Navigators uses seven financial performance metrics to compute a charity's financial score. Each performance metric is measured by a score on a scale of 0 to 10. All the scores and 30 points are added up so that that top possible score is 100 points. The financial metrics are

obtained from the Form 990 that charities file with the IRS and include four measures of financial efficiency (PM1-PM4) and three measures of financial capacity (PM5-PM7):

- PM1: Program Expense Percentage
- PM2: Administrative Expense Percentage
- PM3: Fundraising Expense Percentage
- PM4: Fundraising Efficiency
- PM5: Program Expenses Growth
- PM6: Working Capital Ratio
- PM7: Liabilities to Assets Ratio

When evaluating a charity, the CN assigns the scores on each of the measures according to its financial score conversion and tables, which are available to the public (Charity Navigator, 2016a). Charity Navigator explains the conversion system by the need to recognize operational differences across different types of charities. Before assigning a score to financial efficiency metrics, Charity Navigator also claims that it makes joint cost allocation and indirect cost allocation adjustments. At the same time, nothing is mentioned about adjustments related to the valuation of in-kind donations.

The CN calculates a charity's accountability and transparency score by evaluating the charity against its 20 performance metrics using the data from its Form 990 and website (Charity Navigator, 2020b). The performance metrics are based on a set of good governance practices, policies, and reporting requirements. A charity's score is calculated by subtracting a certain amount of points from the base score of 100 for each performance metric that the charity does not meet according to the CN's table (Charity Navigator, 2020b).

On face value, the CN report cards appear to rely on a more comprehensive set of performance measures than the previously discussed two rating systems while also offering a set of user-convenient composite measures of overall organizational performance as well as its dimensions. At the same time, reliance on the form 990 data and, hence, arguably an oversimplistic treatment of nonprofit performance remain the main weaknesses of the CN evaluations. To address this issue, the Charity Navigator includes descriptive impact information to its report cards by sourcing it from partner services, including GuideStar, ImpactMatters, GlobalGiving, and Classy. This outsourced information, however, does not impact the CN ratings.

ImpactMatters

A startup nonprofit rating agency, Impact Matters, emerged in 2017, aiming to improve nonprofit accountability and donor decision-making by calculating and reporting organizations' impact and cost-effectiveness (ImpactMatters, 2017). At the outset, the agency started providing two services focusing on “service delivery” nonprofits – guided impact reporting (extracting self-reported data estimates of the cost-effectiveness from nonprofits) and nonprofit impact audits (an independent assessments of cost-effectiveness). At the end of 2019, ImpactMatters announced the start of its rating service.

Currently, the agency reports on 1,080 nonprofits. Its report card includes an overall star-impact rating on a five-point scale, an estimate of the charity's cost-effectiveness, and a governance check as shown in the figure 2.4. below.

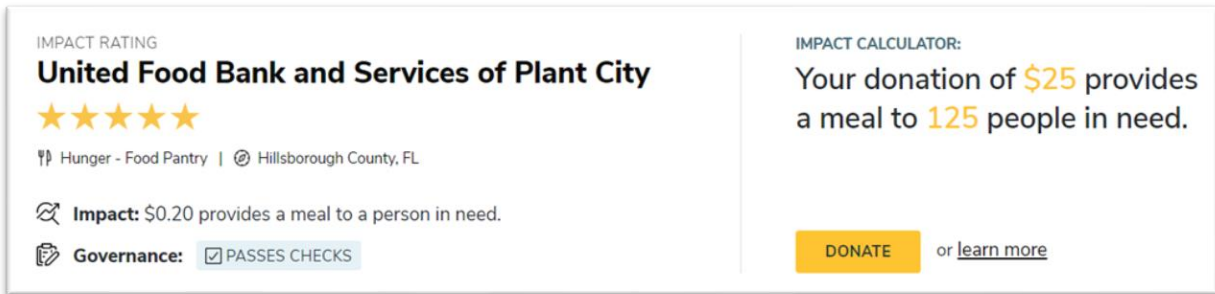


Figure 2. 4: Summary of a report card from Impact Matters. Retrieved May 31, 2020 from <https://www.impactmatters.org/ratings/?q=United+Food+Bank+and+Services+of+Plant+City>. Screenshot by author.

The overall rating is assigned based on the following criteria:

- 5 stars:** The rated program is highly cost-effective.
- 4 stars:** The rated program is cost-effective.
- 3 stars:** The rated program does not meet ImpactMatters' benchmark for cost-effectiveness.
- 2 stars:** After being given an opportunity, the nonprofit chose not to publish impact information.
- 1 star:** There are indications of governance or financial health issues at the nonprofit.

In summary, ImpactMatters takes a different rating approach compared to the other three agencies by attempting to fill the information gap their report cards have been most criticized for. Among the potential shortcomings of the impact-based ratings could be their limited comparability across causes, complexity of evaluation, susceptibility to error, dependence on the available impact-related information in public sources, higher cost, and limited evaluation capacity and pool of nonprofits to choose from (ImpactMatters, 2020a, 2020b, 2020c).

2.3 Content and Comparability of Charity Ratings: Hypotheses

Third-party raters analyze information known to private parties and reveal it to other interested uninformed parties. Typically, they obtain one summary score that aggregates a certain amount of performance information to accomplish this task (Scanlon, Chernew, Sheffler, & Fendrick, 1998). Such information reflects various dimensions of performance, which are measured, transformed into index scores, and assigned some normative values – ratings. Consequently, raters are supposed to make many important decisions as to the content of their ratings, including data sources, sampling, selection of measures, and methodology of computing ratings. As a result, different rating systems might reveal different amounts of information, be driven by different factors, and, more importantly, be ultimately in disagreement with each other. Scanlon et al. (1998) noted that “such disagreements may undermine the public’s confidence in these instruments” (p.13) and cause underutilization of such systems by their potential users in their decision making.

Some research supports these assumptions. For instance, Lizzeri (1999) studied the extent of information revelation and strategic manipulation by quality certification intermediaries. He argued that a monopoly certifier is motivated to reveal only a minimal amount of information by providing a simple pass/fail certificate based on a minimum quality standard. However, as the number of intermediaries grows, competition between them leads to full information revelation. Interestingly, the oldest charity rating agency, the BBB Wise Giving Alliance, uses a very similar rating scale – Meets Standards/Standards Not Met (Gordon et al., 2009). Later entrants to the market of rating charities - the American Institute of Philanthropy, or the Charity Watch, offer progressively more elaborate rating scales.

Zhe Jin et al. (2010) conducted an experimental study of the informational role of certification intermediaries in the context of sportscard grading markets. Consistent with the points made above, they found that the first professional certifier provides less information for uninformed parties than new entrants who differentiate from it by offering finer grading approaches and more precise signals about quality.

Pope (2009) studied patient response to US News and World Report (USNWR) hospital quality rankings. USNWR claimed that it ranked 2000 eligible hospitals based on: (1) a survey of physicians, (2) the hospital-specialty's mortality rate, and (3) a combination of other hospital characteristics. The methodology stated that each factor contributed one-third to the final score. However, the author showed that statistically the reputation score explained over 95 percent of the variation in the score almost entirely driving the rankings, and mortality rates accounted for less than one percent. In that regard, they concluded that "reputation scores" (which are much more variable than risk-adjusted mortality rates) represent more of the final score than the claim of one-third. Thus, the continuous quality score that is provided for each hospital can be essentially thought of as an affine transformation of the reputation score." (p.1156).

The largest rater, Charity Navigator writes that it calculates an organization's final score based on two factors – financial health and accountability/transparency – and that seven financial health indicators contribute equally to the financial health score. Hence, it is reasonable to expect that:

H1: The accountability score explains a substantial portion of the variation in the final score assigned to a charity

H2: Program expenses, administrative expenses, fundraising expenses, fundraising efficiency, primary revenue growth, program expenses growth, and working capital ratio contribute equally to the final score assigned to a charity.

Scanlon et al. (1998) present another useful piece of research underscoring the importance of studying the information role and comparability of performance rating agencies. This research team examined health plan ratings and rating consistency across different plans. They discovered that although on the whole plan ratings were positively correlated, the extent of agreement varied substantially. They write that “the correlations in scores were often weak [and] in several cases there was dramatic disagreement among report cards” (p. 13). Such disagreement among report cards, in turn, might send mixed signals to uninformed parties and undermine the confidence in the instrument. Charity ratings would be most useful and effective if ratings assigned by different agencies did not contradict each other:

H3: Charity ratings issued by major rating agencies will be highly correlated and consistent

Overall, the reviewed literature on the comparability of external quality/performance certifiers admits that we still know little about the behavior of professional certifiers, and few studies have compared external raters (Scanlon et al., 1998; Zhe Jin et al., 2010). Even less is known about the behavior of charity raters. Except for the two descriptive studies by Lowell et al. (2005) and National Council of Nonprofit Associations and the National Human Services Assembly (2005), there is no scholarly research comparing charity rating agencies.

2.4 Data and Methodology

This analysis focuses on the whole population of the charities rated by Charity Navigator (CN) for the fiscal years 2000-2018. These data are available in open access at the Charity Navigator website, and the dataset was obtained using automatic data extraction techniques that involved data scraping directly from the website using R statistical software. The resulting dataset overall contains 102534 observations for 8640 charities rated based on financial data for FYE 2000-2018. The distribution of the ratings based on fiscal years (FY) is presented in the Appendix A.

In addition to the CN data, the analysis will also utilize the American Institute of Philanthropy's Charity Watch (CW) ratings to answer the questions related to the comparability of charity ratings. Access to all CW ratings is provided to paid CW members only through the rater's password-protected website. The dataset that includes ratings on 595 charities obtained from the source website after purchasing a membership by using a similar set of web scraping techniques. Unlike Charity Navigator that provides historical data on its ratings, the Charity Watch makes available only its most recent ratings. The obtained dataset contained ratings assigned to charities based on their financial data for fiscal years 2012 – 2018 with the overwhelming majority of ratings based on the FYE 2016 – 2018 financial reports. The distribution of the CW sample across fiscal years is presented in the Appendix A.

The CN and CW are the two largest quality certifiers that offer some of the most elaborate report cards and ratings with multiple point scales. Both heavily rely on forms 990 for performance information. Both provide their rating data on their websites with names and unique identification numbers of the charities they rate and description of their rating methodologies.

This research focuses on Charity Navigator ratings to address the first question regarding the extent to which charity ratings reveal information to their users. To obtain star ratings, the CN computes continuous scores on each of the following composite dimensions of charity performance: accountability and transparency scores, financial scores, and overall scores. Charity Navigator claims that the financial score is determined as an additive index of 7 measures equally determining a charity's financial score and rating.

To see what performance dimensions and measures represent the variation in the final score and, thus, drive the ratings, this analysis followed the approach Pope (2008) took for determining the drivers of hospital rankings. The continuous performance score is regressed first on all of the variables described by CN methodology and on each component separately. In addition, given there are eight potential drivers of the final ratings in the methodology of Charity Navigator, this analysis conducts a hierarchical linear regression by successively adding more predictors to the model. The statistics of interest in this analysis is the coefficient of determination (R-squared), which shows how much variation in different variables – components of the overall score contribute to the variation in final scores and charity ratings (the technique is identical to the forward model selection based on R-squared). This analysis would allow a comparison of the contribution of different determinants of the composite scores to the rater's claims regarding the content of the summary grades.

Two additional analyses will answer the second question regarding the consistency of different rating systems. One approach used by Scanlon et al. (1998) to compare health plan ratings is to compute the Spearman rank correlation coefficient for the examined rating systems. The magnitude of the estimated correlation coefficient will indicate the extent of agreement among the raters, and the ratings are hypothesized to be highly correlated. Zhe Jin et al. (2010),

however, argued that, because different ratings (grades) are ordinal and due to different grading cutoffs are not readily comparable, computing the raw rank correlation is not a very robust approach. Therefore, an additional analysis in this section will follow the method Zhe Jin et al. (2010) adopted to compare alternative certifiers in the market of sportscard grading.

In the context of charity rating, this analysis uses a sample of 210 charities that is an intersection of the two rating data sets – Charity Navigator and the Charity Watch. This analysis will examine if the two raters agree on the relative performance of any two charities (A and B for further convenience) selected from the sample. The two raters, the CN and CW, will be defined as *strongly consistent* if they agree that the performance of the charity A is superior or equal to that of charity B ($p_A \geq p_B$). If one of the raters decided that $p_A > p_B$, but the other rated $p_A < p_B$, then the two are *strongly inconsistent*. The final alternative, when one of the raters decided that $p_A > p_B$ but the other rated $p_A = p_B$, then the two raters are *weakly inconsistent* for this pair of charities. Such a comparison will be made for all distinct pairs of n charities (the total number of pairs can be calculated as $n!/2(n-2)!$). The results of all the comparisons will be recorded and percentages in which the raters are strongly consistent, strongly inconsistent, and weakly inconsistent will be calculated. This analysis will provide an informative description of the degree of consistency among the two raters.

2.5 Findings

Because this study examines how much each component in the structure of the CN rating contributes to explaining the variation in the rating grades, the available data were further restricted to the observations that included the *Accountability and Transparency* rating, which

Charity Navigator introduced in September 2011 (Charity Navigator, 2016b). As a result, the restricted dataset that included ratings using the CN methodologies 2.0 and 2.1 contained 69,409 observations for 8640 charitable agencies. After recalculating all the Overall Scores from the Financial score and Accountability and Transparency score using the formula that Charity Navigator uses ($Overall\ Score = 100 - \sqrt{\frac{(100 - Financial\ Score)^2 + (100 - A\&T\ Score)^2}{2}}$), the analysis found that Overall Scores in 1018 observations published for 543 agencies did not match the recalculated scores (Appendix A, Figure A.5). Along with observations with missing ratings, the total number of 1093 observations (1.6 percent of all available observations) were also removed from further analysis.

According to the CN methodology (Charity Navigator, 2020a), the *Overall Score* is obtained using a nonlinear, but identical for the two components transformation of the *Financial Score* and *Accountability and Transparency Score*, so the R^2 from the linear model that includes both variables will be less than 100 percent, but the contributions of each component can be estimated relative to it. The results are presented in Table 2.5. They show that, when the R^2 with both components of the score equals 96 percent, the *Financial Score* alone explains 58 percent in the variation, and the *Accountability and Transparency Score* alone explains 51 percent of the variation in the *Overall Score*. Out of the total explained variation, the contribution of the Financial Score into the overall measure appears to be larger by only 20.8 percentage points, and it can be concluded that both measures have substantial and comparable influence over the composite score. This finding is consistent with the first hypothesis that the accountability score explains a substantial portion of the variation in the final score assigned to a charity.

Table 2. 5: Relative contributions of the Financial Score and Accountability and Transparency Score in explaining the variation in the Overall Score

	Model	Adjusted R²	Improvement in R²	Contribution (percent)
1	Overall Score ~ Financial Score	0.58		60.4
	Overall Score ~ A&T Score	0.51		53.13
2	Overall Score ~ Financial Score + A&T Score	0.96	0.38	100

Table 2.6 presents the results of conducting the hierarchical linear regression for the *Financial Score* using a stepwise adjusted R²-based forward selection. The first step regresses the *Financial Score* on each of the components alone. The adjusted R²s are recorded and compared to find the largest one to select the model for the next step, and the procedure is repeated until reaching the full model that includes all the components of the *Financial Score*. The predictors that make the largest increments in adjusted R² in each step are highlighted in the table.

The first interesting result that follows from this analysis is that the full model, despite being additive, with all the predictors included, explains only 54 percent of the variation in the Financial Score. The remaining variation in the Financial Score thus can probably be explained by the normative conversion schemes and adjustments that Charity Navigator applies to the raw financial measures to obtain the converted scores that then added up to convert to a 100-point scale. Thus, it would be fair to conclude that the Financial Score is only partially (54 percent) objective.

Table 2. 6: Hierarchical linear regression using R²-based forward model selection

Step	Model	Adj. R ²	%
1	Financial Rating ~ Program Expenses	0.33	61.1
	Financial Rating ~ Administrative Expenses	0.14	
	Financial Rating ~ Fundraising Expenses	0.22	
	Financial Rating ~ Fundraising Efficiency	0.25	
	Financial Rating ~ Program Expenses Growth	0.18	
	Financial Rating ~ Working Capital (WC) Ratio	0.01	
	Financial Rating ~ Liabilities to Assets (LA) Ratio	0.03	
2	Financial Rating ~ Progr. Expenses + Admin. Expenses	0.34	83.3
	Financial Rating ~ Progr. Expenses + Fundr. Expenses	0.34	
	Financial Rating ~ Progr. Expenses + Fundr. Efficiency	0.37	
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth	0.45	
	Financial Rating ~ Progr. Expenses + Working Capital Ratio	0.36	
	Financial Rating ~ Progr. Expenses + LA Ratio	0.36	
3	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Admin. Expenses	0.46	90.7
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Fundr. Expenses	0.46	
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Fundr. Efficiency	0.49	
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + WC Ratio	0.48	
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + LA Ratio	0.48	
4	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Fundr. Efficiency + Admin. Exp.	0.49	96.2
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Fundr. Efficiency + Fundr. Exp.	0.49	
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Fundr. Efficiency + WC Ratio	0.52	
	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Fundr. Efficiency + LA Ratio	0.52	
5	Financial Rating ~ Progr. Expenses + Progr. Expenses Growth + Fundr. Efficiency + WC Ratio + LA Ratio	0.54	100
	Financial Rating ~ Admin. Expenses + Fundr. Expenses + Progr. Expenses Growth + Fundr. Efficiency + WC Ratio + LA Ratio	0.54	
	Financial Rating ~ Progr. Expenses + Fundr. Expenses + Progr. Expenses Growth + Fundr. Efficiency + WC Ratio + LA Ratio	0.54	

Table 2. 7: Relative contributions of the reported Program Expense Ratio and Fundraising Efficiency in explaining the variation of the Charity Watch grades

	Model	Adjusted R ²	Improvement in R ²	Contribution (percent)
1	Overall Score ~ Fundr. Efficiency	0.34		97.14
	Overall Score ~ Progr. Expenses	0.27		77.14
2	Grade ~ Progr. Expenses + Fundr. Efficiency	0.35	0.01	100

The next interesting finding is that two variables - *Program Expenses* and *Program Expenses Growth* are the major drivers of the variation in the *Overall Score* as the two variables jointly account for 83.3 percent of the variation explained by the model with all predictors

included. Along with the *Fundraising Efficiency*, the three efficiency measures explain over 90 percent of the variation in the *Financial Score*. Also, when the *Program Expenses* variable is included in the model as an explanatory variable, the *Administrative Expenses* and *Fundraising Expenses* do not add any explanatory power to the set of predictors, which is consistent with the fact that the *Administrative Expenses* variable is a linear combination of the former two measures. Overall, this analysis disconfirms the second hypothesis that the seven financial measures that the rating agency uses in its calculation of the *Financial Score* equally contribute to the final score assigned to a charity.

The first step in conducting the consistency analysis for the two charity raters was converting charity ratings assigned by each rating agency to a numeric scale. Charity Navigator rates nonprofits on a five-point scale from zero to four stars. The numeric values were assigned accordingly in the range between zero and four. Charity Watch's grading scale is different from the one used by Charity Navigator. It uses 11 letter-grades from the lowest "F" grade to the highest "A" grade. The letter-based performance grades by the CW were converted into numeric grades using two approaches. First, letters were converted to 11 numeric grades from 0 – 10 to preserve the native CW scale. The second conversion adapted the CW scale to the CN scale converting the 11-point letter scale into a five-point numeric scale to make it similar to the CN scale. The distribution of the original grades and the converting schemes are provided in Figures A1-A4 and Table A1 of the Appendix A. Analyses of consistency were conducted for both converted scales. Consistency analyses could be conducted only for the ratings assigned to the same charity for the same fiscal year. Therefore, the two datasets were intersected based on those two variables. After the three matching nonprofits that had split CN ratings were removed, the

final sample contained 210 matching charities that yielded 21,945 unique pairs of ratings for consistency analysis.

Table 2.8 below shows that the association between the CN and CW score is moderately strong regardless of whether the native or converted numeric scale is used for the CW ratings. In fact, the Pearson correlation coefficient is even somewhat higher when the CW rating is measured on the native 11-point scale.

Table 2. 8: Pearson correlation coefficients between the CN and CW performance grades.

	CN grade ~ CW Grade (on the native 0-10 scale)	CN grade ~ CW Grade (on the adapted 0-4 scale)
Pearson Correlation Coefficient	0.54	0.51

When the rating grades assigned by the two alternative agencies are compared on their native measuring scales, the consistency analysis, results of which are presented in Figure 2.5 and Table 2.9, shows that grades for only 50.2 percent of all distinct pairs of charities in the sample are strongly consistent according to the definition. Another 35.6 percent of grade comparisons in the sample show weak inconsistency in the assigned overall performance grades, whereas 14.2 percent of the compared grades fall in the category of strongly inconsistent. Converting the CW 11-point letter scale to a five-point numeric scale that is consistent with the CN grading scale lead to increased distances between CW’s grades could eliminate differences in the CW grades for some of the compared charities in the sample. This, in turn, would lead to an improvement in inter-rater consistency. The second column in Table 2.9 shows that the overall consistency only slightly improved as the percentage of strongly inconsistent grades decreased by 2.4 percentage points, almost entirely moving to the weakly inconsistent category.

The percent of strongly consistent grades improved by only 0.1 percent, leaving the consistency rate at 50.3 percent.

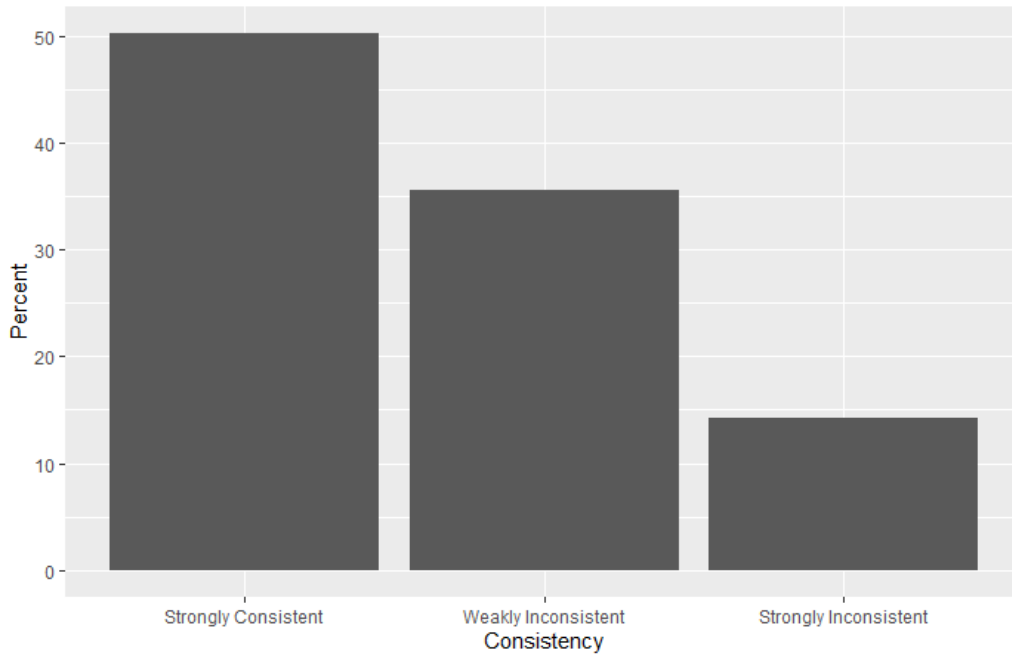


Figure 2. 5: Inter-rater consistency in charity performance grades

Table 2. 9: Cross-rater consistency in rating grades

	Consistency (Percent) when CW Grade on the Native 0-10 Scale	Consistency (Percent) when CW Grade on the Native 0-10 Scale and observations with errors in the CN Overall Rating removed	Consistency (Percent) when CW Grade on the Converted 0-4 Scale
Strongly Consistent	50.2	44.6	50.3
Weakly Inconsistent	35.6	39.5	37.9
Strongly Inconsistent	14.2	15.9	11.8

2.6 Conclusions and Implications

Modern information and computer technologies have created favorable conditions for the charity rating industry, but from the capacity perspective, it still appears to be in its nascent state. As follows from this study, not only the major charity rating agencies rate a relatively small fraction of the reporting to the IRS nonprofit agencies, the assigned performance grades are mainly driven by a rather limited set of measures, even if it looks different on the surface.

This analysis focuses on the informational content and consistency of the major charity raters that rely on the same data source – Form 990. In the case with Charity Navigator, the largest issuer of charity ratings that evaluates and assigns performance grade to over 9000 public charities, the two components that make up the overall performance score are the *Financial Score* and the *Accountability and Transparency Score* with the former contributing 60.4 percent and the latter 39.6 percent to their joint explanatory power. Out of seven predictors of the *Financial Score*, two (*Administrative Expenses* and *Fundraising Expenses*) appear to add no informational content to the model and are redundant. Another pair of measures jointly adds only five percent to the informational capacity of the model. As a result, the financial score is mainly driven by three efficiency measures, which collectively focus on charity program spending levels. In other words, the informational content of the charity performance ratings is lower than it appears on face value. By contrast to the CN, the grades provided by the Charity Navigator are almost entirely driven by the measure of fundraising efficiency, while the two measures used in the calculation of the rating explain only 35 percent of the variation in the assigned grades.

The analysis of inter-rater consistency conducted in this study also shows that there might be a variation in the signals that different rating agencies send to donors about the same charities at the same point in time based on the same information sources. From the perspective of a

potential donor, this may make the process of information search about charitable performance costlier as it requires considering alternative options, learning about details and differences, and choosing among raters. Both low informational content and a too low level of consistency among evaluations provided by alternative raters can also have negative impacts on public trust in charity performance monitoring systems.

CHAPTER 3: INDIVIDUAL DONOR RESPONSE TO CHARITY RATINGS

3.1 Introduction

According to long-established theories, the prohibition from distributing the operational surplus to owners makes private nonprofit organizations a trustworthy alternative to opportunistic businesses – in other words, a vehicle to overcome the contract failure⁵ (Steinberg, 2006). Because of the non-distribution constraint plus entrepreneurial sorting⁶ (Young, 2013), those in control of an organization will be less motivated to compromise on quality or quantity. However, as in traditional market exchange relations where information asymmetry between buyers and sellers about quality leads to market inefficiencies (Akerlof, 1970), information asymmetry concerning organizational performance allows low-quality nonprofit institutions to attract donor resources for unproductive and sometimes even not well-intended uses (Charity Watch, 2018; Kelly, 1998; Salamon, 2012). Repeated high-profile reputational failures involving ineffectiveness, fraud, wastefulness, and lavish spending on executive perks (Attkisson, 2009; Goldberg, 2015; Hoffman, 2006) and the growing negative perceptions about nonprofits undermine public confidence in and future support of these institutions (Interactive, 2006; Kelly, 1998; Light, 2008; Peng, Kim, & Deat, 2019; Rhode & Packel, 2009; Salamon, 2012). As it becomes increasingly evident to the public that the nonprofit status does not prevent organizational leaders from pursuing selfish ends or running inefficient operations, public disenchantment with the third sector leads to questioning the rationale behind nonprofit tax

⁵ According to Hansmann's theory of the nonprofit enterprise (Hansmann, 1980) when "the quantity or quality of service cannot be verified, markets take advantage of informational asymmetries" (Steinberg, 2006, p. 119)

⁶ According to (Young, 2013), "entrepreneurs of different motivations and styles sort themselves out by industries and economic sectors in a way that matches the preferences of these entrepreneurs for wealth, power, intellectual or moral purposes, and other goals with the opportunities for achieving these goals in different parts of the economy" (p.3) and "participants in nonprofit agencies tend to have personal goals and attitudes more consistent with maintaining the quality and integrity of services than do participants in other sectors" (p. 128)

privileges and the adequacy of nonprofit accountability (Herzlinger, 1995; Hoffman, 2006; Kelly, 1998; Salamon, 2012; The Washington Post, 2018).

While government has limited capacity to protect public interest by enforcing nonprofit fiduciary duties (Gilkeson, 2006), making performance information available to the public could facilitate accountability and establish a basis for trust in nonprofit institutions without turning to intrusive regulatory methods (Moxham, 2009; Salamon, 2002). Such a decentralized approach requires sophisticated performance measurement and impact evaluation, which is challenging to implement in the nonprofit practice, and few nonprofits actually use it (Brody, 2002; Ebrahim & Rangan, 2010; Lampkin et al., 2007; Lynch-Cerullo & Cooney, 2011; Moxham, 2009; Rowe, 2012). Furthermore, most individual donors have limited ability to process complex performance reports (Gormley & Weimer, 1999; Lampkin et al., 2007). As a result, information asymmetry remains between a nonprofit and its donors regardless of its use of performance measurement.

Mechanisms that could effectively facilitate performance-based accountability in the nonprofit sector thus should satisfy demands that go beyond traditional performance measurement. Besides valid and comprehensive analysis, they must provide independent, objective, and regular assessment. In addition to that, information must be relevant, comprehensible, easily accessible to nonprofit donors. In theory, systems that have potential to accommodate such conflicting demands are known as organizational performance report cards (Gormley & Weimer, 1999). Some researchers have argued that nonprofit performance report cards (typically known as nonprofit watchdog groups or charity ratings) offer excellent performance standards for advising donors and may become a potentially powerful monitoring instrument to address accountability and performance concerns (Gilkeson, 2006; Herzlinger, 1995). At the same time, the tool can only be effective if intended users meaningfully and

consequentially refer to the metrics embedded in charity ratings to inform their donative decision making. Nonetheless, the scholarly literature does not agree the regulatory effectiveness of nonprofit rating systems.

This study examines the effects of performance report cards on individual behavior. In particular, since performance information is subject to biased interpretation by individuals (Bækgaard & Serritzlew, 2015), this study draws on the model of the perceptual determinants of donor behavior (Sargeant, Ford, & West, 2006), the theories of the nonprofit supply, and the broader literature on nonprofit performance measurement and performance report cards to examine how measures presented in charity ratings affect individual giving decisions. It also explores the mediating role of donor perceptions of nonprofit performance and donor trust in nonprofit organizations as well as the moderating effects of certain donor characteristics. The research simultaneously focuses on two salient measures that rating agencies and the broader public heavily rely upon – a charity’s overhead spending ratio and its composite performance score. Additionally, it investigates a potential interaction between them. By reporting the findings from a randomized survey experiment that recreated a realistic decision-making situation, this study extends our limited scholarly understanding of individual reactions to nonprofit performance measures, the mechanisms facilitating them, and, therefore, the regulatory potential of publicized performance grades. Furthermore, the results of this research suggest significant practical implications for future performance monitoring policies and measurement practices, including the content, design, and use of nonprofit performance report cards.

The next section overviews the theoretical and empirical literature on the relationship between nonprofit performance and individual charitable giving. The following section formulates several testable hypotheses about how donors respond to performance measures in

charity ratings. The Methodology and Measurements sections describe the experiment to test the hypotheses. The final section presents and discusses the findings, directions for further research, and limitations.

3.2 Performance, Trust, & Individual Donor Behavior

A wide variety of extrinsic and intrinsic factors drive individual giving behavior. The classic theories focus on public benefits, private benefits, and the price of charitable giving (E. Brown & Slivinski, 2006; Vesterlund, 2006). Public benefits are altruistic, driven by a desire to create common goods and see the needs of others fulfilled (e.g., caring about others). Private benefits accrue to the donor. They include “warm glow” or feeling good about making a donation, prestige, self-esteem, recognition, avoiding guilt, or scorn, etc. Altruistic donors probably care more about the quantity and quality of services provided, which would drive donors’ concerns about organizational performance and influence giving decisions. But E. Brown and Slivinski (2006) write that even “warm-glow motive [is] centered on inducing output rather than simply donating dollars” (p. 145). Similarly, deriving good reputation is more likely by supporting an organization with a good performance record rather than one with a poor standing.

Hirschman (1970) provides a useful conceptual framework that describes the behavior of the customer facing a decline in an organization. According to the theory, when absolute or relative quality of a provided product declines, the dissatisfied customer has only two options – economic (“exit”) or political (“voice”). Exit implies the customer’s withdrawal from the relationship. Voice is an attempt to actively change the organization’s practices. Voice is a relatively costly, and, unless exit is unavailable or the individual is a member of the organization,

exit is the prevalent reaction. The framework also categorizes customers as alert and inert with respect to quality. If an organization has a mix of alert and inert customers, revenue, as Hirschman (1970) explains, “will normally decline steadily as quality drops” (p. 23) without causing too much damage that would lead to the firm’s immediate failure.

However, when deep information asymmetry is present, as in the context of most public charities, individual donors cannot observe and compare the quantity or quality of services provided. Therefore, Sargeant et al. (2006) emphasize the importance of donors’ perceptions of the organization, its output, management, performance, and various benefits they might derive by supporting a nonprofit. Perceptual factors can also be divided into perceptions of private and public benefits from making a donation (Sargeant et al., 2006).

Drawing on social exchange theory, Sargeant et al. (2006) further distinguished three categories of perceptual benefits: demonstrable, emotional, and familial. Demonstrable benefits refer to selfish economic considerations, such as perceptions of one’s improved standing in the donor’s social group and may result from the visibility of giving. Emotional and familial benefits are associated with donors’ emotional experiences. Their argument states that a charitable act can evoke positive emotions, desirable mood changes, or good feelings, and might be an indication of donor commitment to a particular cause. Sargeant et al. (2006) did not find evidence supporting the demonstrable benefits argument, but they found emotional and familial benefits to be significant and direct (bypassing trust) drivers of individual willingness to donate.

Potential donors use various information cues to shape their beliefs about how a nonprofit will use a charitable gift and fulfill its fiduciary obligations. Such beliefs are viewed as trust (in a

specific nonprofit organization), and this construct mediates the relationship between perceived performance and giving behavior, as the Figure 3.1 below shows.

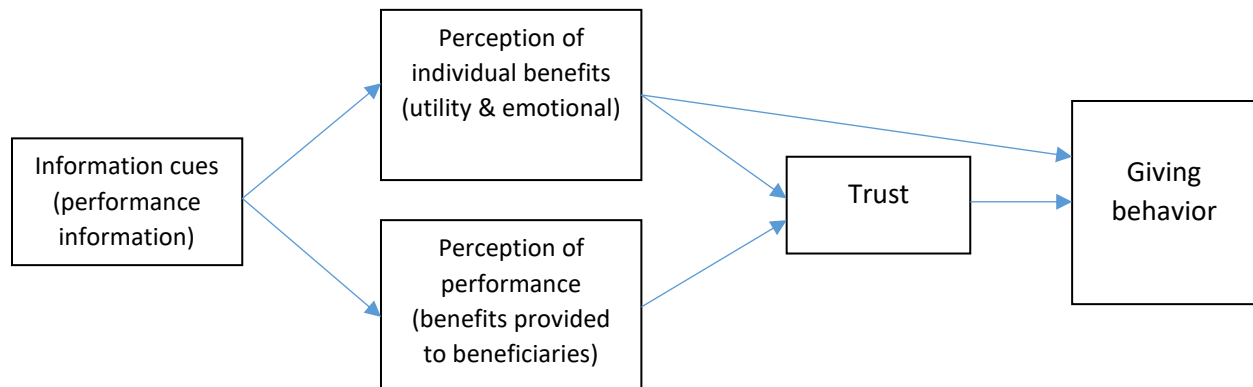


Figure 3. 1: Perceptual Determinants of Charitable Giving

Trust and perceptions of performance drive donor decisions. Individuals are concerned about the extent to which nonprofits help people, spend money wisely, behave ethically, and run programs well (Interactive, 2006; Light, 2008). Positively perceiving a charity’s performance increases trust in that organization (Sargeant et al., 2006) Individuals’ perceptions of higher quality of information about how donor money is used can impact giving via trust (Sargeant et al., 2006). In one 2013 survey, 81 percent of respondents saw impact as the most essential factor in deciding whether to donate, 75 percent looked for online information about nonprofits, 57 percent made a donation after watching an online video, and 47 percent researched across nonprofits before donating (Google, 2013).

According to the research focusing on nonprofit efficiency, donors incorporate performance concerns in their decision-making. Donor perceptions of nonprofit efficiency influence both their evaluation of an organization and their propensity to give (Bennett & Savani,

2003). Private donors are particularly sensitive to organizational overhead costs – the share of income spent on administration and fundraising. A 2008 survey on charitable confidence⁷ revealed that a majority of Americans believe that charities waste money as well as serious public concerns about nonprofit spending and inefficiency even among strong charity supporters (Light, 2008). The majority of people are concerned about how much charities spend on administration and marketing and think that most charities spend more than they should (Bennett & Savani, 2003). Charities that spent more of their donor contributions directly on programmatic activities and actively informed the public about this were more successful in attracting donations. They wrote that “value for money was cited more frequently as a factor in choosing charities than were specific charitable objectives” (p. 328), implying crucial importance to donors of the quality of a charity in terms of its efficiency and effectiveness. More recent research provides some experimental evidence indicating the unpopularity of overhead costs among donors. For instance, Gneezy, Keenan, and Gneezy (2014) concluded that nonprofit supporters tend to avoid organizations with high overhead ratios, which the authors argued hurts organizational outcomes. Other experimental findings suggest a more complex relationship. Information that a charity is efficient leads some donors to give more since their money creates more value, but leads others to give less since they can achieve the same value at a lower cost (Butera & Horn, 2014).

Studies of e-philanthropy also indicate that trust has a big influence on donative intentions (Burt & Dunham, 2009). Drawing on personal psychology and e-commerce literature, Burt and Dunham (2009) argued that those with a higher level of dispositional trust

⁷ Conducted for the Organizational Performance Initiative at New York University’s Robert F. Wagner School of Public Service

(trusting others) might be more likely to trust a nonprofit. Transactional trust (a donor's perception of how their donation will be used - performance) can be improved by providing relevant and rich information about the organization and its activities (Burt & Dunham, 2009; Burt & Gibbons, 2011).

Overall, the theory argues that various dimensions of nonprofit performance have strong relevance to donor decision-making. Some nonprofit supporters not only demonstrate a passion for their cause, but also consider an organization's efficacy in carrying out their missions (Hart, Greenfield, & Haji, 2007). Hence, under complete information, organizational performance would be a significant factor stimulating some potential donors to prioritize their donations to the uses that promise the highest value of benefits that accrue to beneficiaries per dollar contributed. In practice, however, potential donors must rely on various information cues such as overhead ratios, annual reports, marketing communications, articles in mass media, and other sources of incomplete and often biased information. Even in the presence of robust and objective performance measurement systems and reports, substantial information asymmetry between nonprofits and their supporters would likely remain due to the cost of processing sophisticated performance evaluations (Lampkin et al., 2007). The following section discusses nonprofit performance ratings as a potentially significant information cue that can shape individual perceptions of charity performance and thus influence their giving decisions.

3.3 Performance Ratings and Donor Reactions: Research Hypotheses

Third-party quality certification services have emerged in recent decades to provide an independent, objective, comprehensive, easy-to-interpret, low-cost-to-access, and convenient

tool to assess nonprofit performance. If individuals care about nonprofit performance in their donative calculus, as the theory argues, and if they regard charity ratings as a valid source of such information, they will use the evaluations provided in those ratings to inform their perceptions of nonprofit organization performance when making giving decisions. Quality ratings have proven to be a useful informational instrument for improving efficiency in many sectors of social-economic activity including debt markets, restaurants, healthcare, and sport cards (Capeci, 1991; Jin & Leslie, 2009; Jin & Whalley, 2007a; Johnson & Kriz, 2002; Luca, 2011; Pope, 2009; Zhe Jin et al., 2010), Charity rating agencies too can serve as quality certification intermediaries that correct resource allocation in the nonprofit sector based on organizational performance. Such rating services could reduce the cost of performance information search and interpretation to a potential donor. This, in turn, would “lubricate” donor decision-making processes and help guide the flow of charitable dollars towards “good” organizations.

Nonprofit scholars started studying the impacts of the third-party nonprofit performance report cards more than a decade ago, but do not agree on whether charity rating systems affect giving decisions or not. The available to date literature is split on the question of whether charity rating systems affect giving decisions or not. Using organization-level data, Chhaochharia and Ghosh (2008) found that the charities that received the lowest ratings from the American Institute of Philanthropy received fewer contributions. The authors concluded that the tool provides informational value to donors and reduces information asymmetry. Sloan (2009) found that New York charities with “pass” grade from the Better Business Bureau’s Wise Giving Alliance received an increase in contributions compared to those that did not have a rating. At the same time, the “did not pass” label did not affect donations. Using a random sample of 405

charities rated by Charity Navigator, Gordon et al. (2009) found evidence that a change in ratings is associated with a corresponding change in donations. Brown, Meer, & Williams (2014) conducted a laboratory experiment to explain charity choice and willingness to donate, where they varied whether nonprofit performance ratings were displayed. They concluded that third-party ratings influence charity choice and suggested that they may also increase donations. Peng et al. (2019), in their experimental study of nonprofit reputation, found that the availability of a third-party accreditation increases contributions.

In contrast to the findings cited above, a few other researchers concluded that charity ratings are irrelevant to donative decision making. Specifically, one of the earliest attempts to evaluate the effects of third-party performance grades on private giving is by Silvergleid (2003), where the author concluded that the AIP grades did not significantly influence donation levels. Using organization-level data for 90 nonprofits in the state of Washington, Szper and Prakash (2011) tested whether charity ratings affected charitable giving and found no evidence supporting the hypothesized relationship. Interestingly, the qualitative analysis they conducted revealed that the sampled charities did not believe that rating information enters donative decision making either. Finally, the results from the most recent experimental study also cast doubt on the signaling effectiveness of nonprofit rating systems. Tremblay-Boire and Prakash (2017), in their study of the effects of charity participation in a voluntary regulatory program, found no evidence that the availability of a three-star grade provided by Charity Navigator influence individual willingness to donate.

Overall, the nonprofit theory expects charity supporters to consider third party performance auditing as a strong information cue about nonprofit organization quality. However, the sum of available empirical evidence can neither confirm nor reject the argument. This

unsatisfactory outcome may have multiple reasons. First, variation in salience and credibility of specific rating systems (e.g., AIP, BBB, Charity Navigator, etc.) over time, research question focus (e.g., charity choice vs. propensity to donate), or research designs (pass/not pass rating vs. no information, average rating vs. no information, third-party accreditation vs. no accreditation) may create different contexts that lead to the inconsistent findings. Second, the analytical methods and data (regression analyses using organizational level data; laboratory experiments; survey experiments) can make a difference too. For instance, Hirschman's (1970) framework well explains why observational model could fail to prove ratings' effectiveness. In particular, Hirschman explains that "no matter what the quality elasticity of demand, exit could fail to cause any revenue loss to the individual firms if the firm acquired new customers as it loses the old ones" (p. 26). This behavior is consistent with the highly-inefficient segment of charities who invest a relatively large proportion of their revenue in fundraising, including through contracting paid solicitors (Kelly, 1998). Also, an experimental study could fail to detect the hypothesized affect if the experimental stimulus is not strong enough. According to Hirschman's (1970) theory, a certain level of deterioration in an organization's service may not be sufficient to trigger a customer's withdrawal.

Because the lack of consensus presented in the literature findings can be context-dependent, this study conducts a focused, in-depth examination of nonprofit performance report cards to clarify the nexus between performance measures embedded in charity ratings and giving allocations. First, it sets to determine whether there is a causal relationship between the key measures embedded in third-party charity ratings and donor perceptions of a rated organization's performance, donor trust, and willingness to donate. Second, the uncertain findings from the extant literature suggest that even if rating information affects donative decisions, those effects

are probably not drastic and may be nonlinear. To capture potentially subtle effects of the rating signaling on individual giving decisions, this study focuses on the extreme values of a rating performance scale. Finally, to gain a more accurate perspective on the ratings' potential to make a difference, this research is focusing on individual willingness to donate to a nonprofit agency captured through conditional giving levels rather than charity choice. Hence, the first set of hypotheses relating star-rating performance cues with individual willingness to give posits that:

H1a: Providing information about a public charity's low overall performance rating will decrease donor perceptions of the charity's performance. Information about a public charity's high overall performance rating will increase donor perceptions of the charity's performance.

H1b: Providing information about a public charity's low (high) overall performance rating will generate a lower (higher) degree of trust in that organization

H1c: Providing information about a public charity's low (high) overall performance rating will lead to a lower (higher) willingness to donate to that charity

Public opinion surveys show that nonprofit efficiency concerns individuals, and a large body of academic literature focuses on issues related to nonprofit overhead spending. A contentious scholarly debate regarding the appropriateness of the overhead cost as a performance measure continues. Bennett and Savani (2003) argued that public reaction to charities' levels of overhead cost has been irrational. Gneezy et al. (2014) wrote that it could hurt nonprofits' ability to fulfill their mission as it creates barriers to investing in nonprofit infrastructure and management capacity. Brooking's report wrote that rating agencies and IRS punish capacity building by using that label (Light, 2008). Although the academic and professional communities

tend to agree on the many shortcomings and side effects of using the overhead cost as a measure of efficiency, it is still broadly used⁸ and may remain a substantial factor in donor decision making (E. Brown & Slivinski, 2006; Gneezy et al., 2014; Light, 2008; Rhode & Packel, 2009; Sargeant et al., 2006). Because individuals interpret performance information through the lens of their preexisting personal beliefs and, in turn, the overhead cost is a measure that individuals easily relate to and may have strong beliefs about, potential donors are expected to respond to the level of nonprofit overhead spending:

H2a: Providing information about a public charity's low (high) overhead cost will increase (decrease) a donor's trust in the charity

H2b: Providing information about a public charity's low (high) overhead cost will generate a higher (lower) degree of trust in that organization

H2c: Providing information about a public charity's low (high) overhead cost will lead to a higher (lower) willingness to donate to that charity

Third-party performance ratings intend to offer more comprehensive and balanced indicators of an organization's quality than any measure such as the overhead ratio alone. Besides a variety of financial health and efficiency ratios, they incorporate measures of transparency, accountability, and governance in their evaluation methodologies and demonstrate attempts to improve their measurement methodologies (Charity Navigator, 2016b). Typically, raters' grades, as composite measures, already incorporate information on a charity's overhead cost. Given this fact, it would be reasonable for users of charity ratings to discount the overhead

⁸ Charity raters typically report overhead ratios along with the composite star- or pass-grades.

indicators entirely while using performance ratings. On the other hand, measures of the overhead cost may seem to be more transparent, relatable, and convincing to individuals. Although the indicators of overhead spending in raters' report cards do not add any additional information to the composite measures, their mere presence on a report card may have significant moderating influence:

H3a: Providing the information about a charity's low (high) overhead cost strengthens (weakens) the effect of its low (high) charity rating on the perceived organizational performance.

H3b: Providing information about a low (high) level of overhead cost moderates the effect of the low (high) charity rating on the level of trust in the nonprofit.

H3c: Providing information about a low (high) level of overhead cost moderates the relationship between an organization's low (high) charity rating and a donor's willingness to donate.

Drawing on social exchange theory, Sargeant et al. (2006) distinguished three categories of perceptual benefits: demonstrable, emotional, and familial. Demonstrable benefits refer to selfish economic considerations, such as perceptions of one's improved standing in the donor's social group and may result from the visibility of giving. Emotional and familial benefits are associated with donors' emotional experiences. Their argument states that a charitable act can evoke positive emotions, desirable mood changes, or good feelings, and might be an indication of donor commitment to a particular cause. Sargeant et al. (2006) did not find evidence supporting the demonstrable benefits argument, but they found emotional and familial benefits to be significant and direct (bypassing trust) drivers of individual willingness to donate.

According to Hirschman's (1970) framework, nonsubstitutability among two products is an important factor that prevents the customer from exit. Considering that nonprofit donors may derive emotional benefits (Sargeant et al., 2006), this suggests that the influence of performance ratings on donative allocations among charities may vary depending on which of the causes under consideration appear to be more emotionally appealing to the donor. Therefore, in a case of deciding between two similar charities, emotional benefits are likely to be similar for the two (close substitutes), and performance ratings should drive the willingness to donate through perceived performance and trust in the organization. If a person cares more about a particular cause or mission – in other words, derives emotional or familial benefits from supporting the cause – then this commitment (mission valence) will affect one's willingness to donate beyond the influence of performance and trust:

H4: Relationship between charity ratings and giving behavior will be stronger when mission valence is weak

Lastly, Grimmelikhuijsen and Meijer (2012) draw on social psychology to argue that preexisting characteristics of people, such as knowledge, beliefs, and attitudes, affect how individuals process and interpret information. Cognitive dissonance theory, as well as the theory of motivated reasoning (Bækgaard & Serritzlew, 2015; Grimmelikhuijsen & Meijer, 2012), argue that individuals interpret new information in ways that confirm their prior beliefs about the world and discount evidence that does not fit their beliefs. Such biased processing means different people will interpret the same information differently. Specifically, individuals with high levels of general trust in nonprofits would be less sensitive to the influence of external performance evaluations than those who are less trusting. At the same time, more altruistic

individuals must care more about nonprofit output, so they should be more sensitive to performance grades:

H5a: The effect of performance rating information will be weaker for individuals with a higher level of general trust in quality of nonprofits

H5b: The effect of performance rating information will be stronger for more altruistic individuals

3.4 Methodology

I use a randomized survey experiment to test the proposed hypotheses. A mixed experimental design employed in this study relies primarily on four conservative between-subject comparisons but also takes advantage of two within-subject measures with controlling for order effects. The experiment was embedded in an online survey and delivered to a sample provided by Qualtrics Panels using the Qualtrics online survey platform.





The experiment randomly assigned participants into four groups. Each group received a performance report card with information describing two of four charities with national or global missions and difficult to measure outcomes⁹. Two of the charities had the lowest (one-star) overall performance rating but a low (1 to 10 percent) overhead spending level, and the other two had the highest (four-star) rating but a relatively high overhead cost (32-35 percent). Each participant received a report card on one Low-Rating-Low-Overhead (LRLO) and one High-Rating-High-Overhead (HRHO) organization. To manipulate the variables of interest, the report

⁹ The charities were selected from the pool of organizations publicly rated by Charity Navigator and, to satisfy the stated criteria, represented medical research and children education policy and relief related causes.

card in each of the experimental groups displayed a different set of performance indicators as information heuristic for subjects to form their perceptions of organizational performance, determine how much they would trust each organization, and choose how to allocate the budget among the two competing agencies. In the no-treatment condition (T1) where the report card included only the charities' names, classification categories, corresponding causes, mission statements, self-described accomplishments, and total revenue level. The second treatment (T2 - Overhead) also included the overhead spending ration but not the performance rating. The third (T3 - Rating only) condition included the base information plus the performance rating, but not the overhead ratio. Finally, the Rating and Overhead (T4) condition displayed both the overhead ratios and the charity ratings in the report card. The experimental conditions are summarized in Table 3.1 (the complete report cards for the four selected charities are presented in Appendix B, Table B.1).

For examining hypothesis H4, the study implements an experimental manipulation of mission valence into the research design. To that end, the four available nonprofit pairs were selected so that in two of them, both charities addressed somewhat similar causes (e.g., medical research). In the other two pairs, the organizations served fundamentally different purposes (e.g., medical research and children education policy).

Table 3. 1: Experimental conditions

T0: No-treatment Condition	T1: Overhead Condition	T2: Rating Condition	T3: Rating & Overhead Condition
<input checked="" type="checkbox"/> Base information <input type="checkbox"/> No Ratings <input type="checkbox"/> No Overhead	<input checked="" type="checkbox"/> Base information <input checked="" type="checkbox"/> Overhead exp. <input type="checkbox"/> No Ratings	<input checked="" type="checkbox"/> Base information <input type="checkbox"/> No Overhead exp. <input checked="" type="checkbox"/> Charity Ratings	<input checked="" type="checkbox"/> Base information <input checked="" type="checkbox"/> Ratings <input checked="" type="checkbox"/> Overhead exp.
Example			
CHILDREN’S RELIEF MISSION Base information	CHILDREN’S RELIEF MISSION Base information + Program expenses: 99.1% Overhead: 0.8%	CHILDREN’S RELIEF MISSION Base information + 	CHILDREN’S RELIEF MISSION Base information + Program expenses: 99.1% Overhead: 0.8% + 
STAND FOR CHILDREN LEADERSHIP CENTER Base information	STAND FOR CHILDREN LEADERSHIP CENTER Base information + Program Expenses: 64.7% Overhead: 35.2%	STAND FOR CHILDREN LEADERSHIP CENTER Base information + 	STAND FOR CHILDREN LEADERSHIP CENTER Base information + Program Expenses: 64.7% Overhead: 35.2% 

Participants were informed that the goal of the survey was to study which public charities individuals trust and feel confident deserve charitable contributions. The instructions told the subjects that the researcher had \$100 to donate to charity and asked the subjects to decide how to allocate the money between the two organizations. Participants were told that the researcher would allocate the \$100 based on their recommendation¹⁰. Then, the experiment proceeded to the section where the subjects were randomized into their experimental conditions and allocated their donations. Lastly, the participants answered a series of questions about their perceptions of

¹⁰ After the completion of the research project, each of the organizations would actually receive the proportional share of the amount based on the average allocations

both charities, their behavioral characteristics (personality trust, altruism), and demographics. The survey also included a set of quality check questions to make sure the survey participants paid attention and meaningfully answered to the questions. The resulting study sample included 873 subjects, and its characteristics, including the break downs by treatment groups, are presented in Appendix B, Table B2.

In summary, the experimental approach allows significant flexibility in meeting research data needs and can deliver exceptional internal validity, including the establishment of causality (Charness, Gneezy, & Kuhn, 2012; James, 2011). A distinctive characteristic of this experimental strategy to further strengthen the internal validity of the findings is that it approximates a realistic decision-making situation when an individual who is asked to make a consequential donative decision is facing a budget-constrained choice among real nonprofits.

3.5 Measurement

The primary outcome of interest in this research is the donation allocation preference (willingness to donate). The behavior was induced and measured by asking the participants to allocate a designated amount of money between two charities after reading their performance report cards. Specifically, the question stated the following: “Please tell us how you would prefer to allocate \$100 to the two charities (you can split the amount in any proportion you want so that the total donation does not exceed \$100)”. The participants entered their dollar allocations into the survey form.

The theoretical argument constructed in this paper also refers to a few intervening and moderating behavioral constructs, including perceived performance, dispositional (personality)

trust, preexisting trust in nonprofits in general, trust in a specific nonprofit organization, emotional and familial benefits. Sargeant et al. (2006) describe trust as “the extent of donor belief that a charity will behave as expected and fulfill its obligations” (p. 2). Burt and Dunham (2009) defines it as “an expectation (a trust) that a donation made to an aid agency for a specific crisis or cause will be used towards that specific crisis or cause” (p. 126). This research will rely on a five-item scale used by both groups of authors to measure trust in a specific nonprofit organization. The question items are listed in Appendix B. Each item in this measure is rated on a 5-point Likert scale where 1 = Strongly Disagree to 5 = Strongly Agree. The trust score is obtained by averaging the scores on each of the items and ranges between 1 and 5. This construct demonstrated a high level of internal consistency ($\alpha = 0.94$). Following the work of Burt and Dunham (2009), dispositional (personality) trust describes one’s “tendency to attribute benevolent intent to others (e.g., to believe that others have good intentions), and suspicion that others are dishonest (e.g., to suspect hidden motives in others—reverse-scored)” (p. 129). Altruism, in turn, is defined as a measure of selflessness and concern for others. Both measures are captured using items from the International Personality Item Pool (2007) (L. R. Goldberg et al., 2006). The measure for trust in nonprofit organizations, in general, is borrowed and adapted¹¹ from Grimmelikhuijsen and Meijer (2012). Finally, perceived performance and utility are measured using multi-item scales from Sargeant et al. (2006) with some modifications appropriate for the context.

¹¹ The original variable measured trust in governments in general

3.6 Findings

3.6.1 Perceptions of Overall Performance

Figure 3.2 and Table 3.2 show that the mean levels of perceived overall nonprofit performance within the reference group (T1) were nearly at the same level for the paired organizations with different overhead ratios and third-party performance grades. In the decision setting where both performance measures of interest were excluded from consideration, both the Low-Rating-Low-Overhead (LRLO) and High-Rating-High-Overhead (HRHO) charities averaged at 3.75 on a five-point scale. In treatment T2, where the overhead spending ratios were embedded into the report cards, the level of confidence in an organization's overall performance increased for the low-overhead (LO) and lowered for the high-overhead (HO) charity. Both changes were statistically significant at the one-percent level. In terms of practical significance, the size of the effects (as measured by Cohen's d) is different for the low- and high-overhead nonprofits. When the overhead is presented, the effect size for an LO entity is 0.47 (moderate) and for the HO entity is -0.3 (rather small). As presented in Table 3.3, the within-treatment difference for the overhead group is highly significant based on a paired t -test, and the effect size is 0.68, which is moderately large according to the normative convention.

In the group where the charities' star-rating was the only additional decision cue added to the report card (T3), the experiment yielded a similar within-group effect size ($d=0.68$). However, this effect is comprised of a highly-significant and moderately-strong ($d=0.54$) decrease in the level of confidence in the performance of the low-rating (LR) charity and a substantially smaller-size increase ($d=0.22$) in confidence for the high-rating (HR) organization, which appears to be significant only at the five-percent level.

Finally, the results for group T4 show the level of perceived performance that is statistically not different from the baseline condition. In other words, the contrasting values of the two performance indicators offset each other: a low overhead ratio remedied the negative effect of the low rating, whereas a high overhead damaged the potential perceptual improvement from the high third-party performance rating.

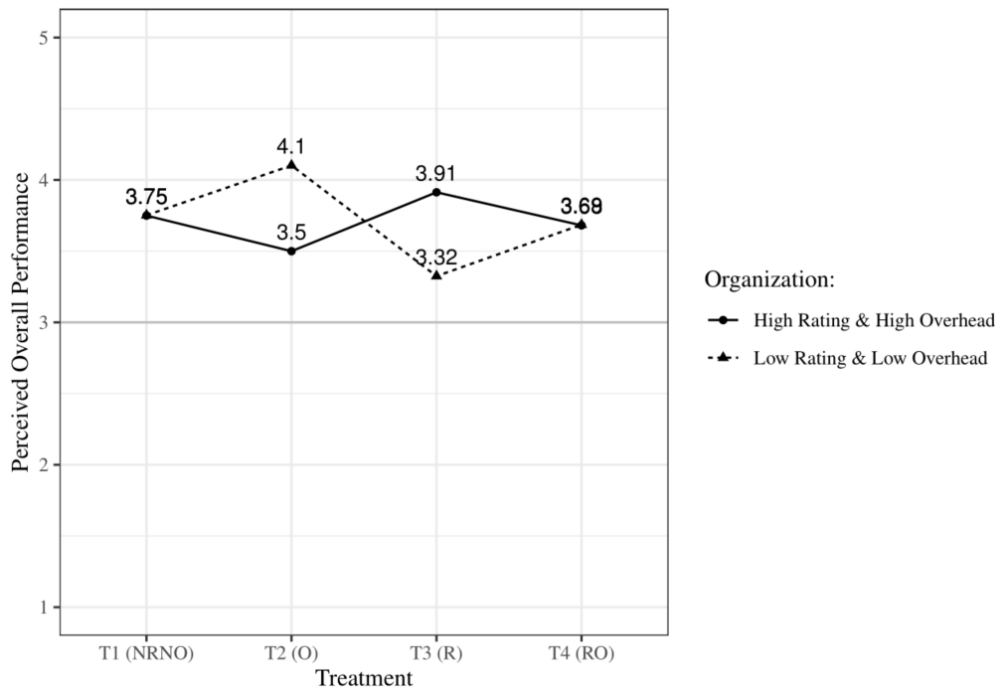


Figure 3. 2: Differences in perceived overall performance across treatments

Table 3. 2: Differences in perceived overall performance across treatments

	Low-Rating-Low-Overhead (LRLO)	High-Rating-High-Overhead (HRHO)
(Intercept)	3.75 *** (0.06)	3.75 *** (0.06)
Treatment T2 (O)	0.35 *** (0.08)	-0.25 ** (0.08)
Treatment T3 (R)	-0.43 *** (0.08)	0.16 * (0.08)
Treatment T4 (RO)	-0.07 (0.08)	-0.07 (0.08)
Observations	873	873
R ² / adjusted R ²	0.107 / 0.104	0.033 / 0.030

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 3. 3: Within-group differences

Treatment	Test statistic	df	P value	Alternative hypothesis	Mean of x
T1	-0.05	207	0.96	Two sided	-0.002
T2	-7.61	224	0.00 * * *	Two sided	-0.602
T3	8.36	213	0.00 * * *	Two sided	0.589
T4	-0.06	225	0.95	Two sided	-0.004

To get a more elaborate picture of the differences in perceptions of nonprofit performance, the following analysis separately examines the two individual performance indicators that comprise the composite performance score - *perceived impact* and *perceived efficiency* spending money - using the ordinal logistic regression analysis.

3.6.2 Perceived Impact

Table 3.4 provides the results of the ordinal logistic regression analysis for the *perceived impact* as the outcome variable. Looking at the cut-points, we can see that the log-odds that individuals in the no-treatment group (T1) express a certain level of agreement (from Strongly Disagree to Strongly Agree) about an organization’s capacity to make an impact are nearly identical for the LRLO and HRHO organizations. The coefficients on T2 show that presenting a low-overhead ratio statistically significantly increases an individual’s propensity to agree with the impact statement. By contrast, the effect of presenting the high-overhead information does not reach statistical significance. According to the estimates for T3, the information about a charity’s low star-rating significantly lowers individual propensity to agree with the impact statement. At the same time, the information about a charity’s high rating does not lead to a significant change in individual perceptions of the organization’s capacity to make an impact. Finally, presenting both the low rating along with low overhead on the report card (T4) has a

significant effect in the same direction as in T3 condition but with a smaller magnitude, thus confirming the moderation effect of the low overhead (the difference T4 - T3 also remains significant). By contrast, presenting the high rating along with a high overhead makes no significant difference in the individual propensity to agree with the impact statement compared to the no-treatment group or rating-only group.

Table 3. 4: Perceived impact across treatments (ordinal logistic regression)

<i>Predictors</i>	LRLO		HRHO	
	<i>Log-Odds</i>	<i>Std. Error</i>	<i>Log-Odds</i>	<i>Std. Error</i>
T2 (O)	0.36 *	0.18	-0.23	0.18
T3 (R)	-0.87 ***	0.18	0.12	0.18
T4 (RO)	-0.45 *	0.18	0.10	0.18
1 2	-4.04 ***	0.25	-4.83 ***	0.40
2 3	-2.48 ***	0.16	-2.85 ***	0.19
3 4	-0.98 ***	0.13	-0.92 ***	0.14
4 5	1.13 ***	0.14	1.10 ***	0.14
Same models with T3 (R) as the reference group:				
T1 (NRNO)	0.87 ***	0.18	-0.12	0.18
T2 (O)	1.23 ***	0.18	-0.35 *	0.18
T4 (RO)	0.42 *	0.18	-0.02	0.18
1 2	-3.17 ***	0.24	-4.95 ***	0.40
2 3	-1.61 ***	0.15	-2.98 ***	0.19
3 4	-0.11	0.13	-1.04 ***	0.13
4 5	2.00 ***	0.15	0.98 ***	0.13
Observations	873		873	
Cox & Snell's R ² / Nagelkerke's R ²	0.058 / 0.063		0.006 / 0.006	

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Figure 3.3 visually demonstrates how the log-odds estimates translate into predicted probabilities for each level of donor confidence in a nonprofit's capacity to make an impact on its cause. The left facet shows the probability changes across the treatment groups for the LRLO condition (all statistically significant effects), and the right panel shows the probabilities for the HRHO entity (insignificant differences).

Elaborating on the insights from the regression output, the left facet shows noticeably broader variations in the predicted probabilities corresponding to each response level across the treatment groups for the LRLO compared to the HRHO entity displayed in the right facet of the figure. For instance, we can see how the inclination to *Strongly Agree* increases while uncertainty (*Neither Agree nor Disagree*) diminishes widening the spread between the two from five to 16.5 percentage points for a LO-charity once the overhead ratio shows up in the report card. Showing the low rating leads to even wider differences across all levels of propensity to agree with the impact statement: the probability of *Strongly Agreeing* drops by 12.5 percentage points from 24.4 percent to 11.9; the probability of *Somewhat Agreeing* drops from 48.4 to 41.0 percent; the probability of *Neither Agreeing nor Disagreeing* increases from 19.4 to 30.4 percent; and the probability of *Somewhat Disagreeing* increases from 6.0 to 12.7 percent. The availability of both performance indicators makes a similar, although weaker, effect to that caused by the low rating only, suggesting a moderation effect. Finally, the right facet shows that the probability changes across the treatments for the HRHO entity are substantially smaller, which indicates that neither a high rating improves nor a high overhead significantly erodes individual perceptions of an organization's capacity to make an impact.

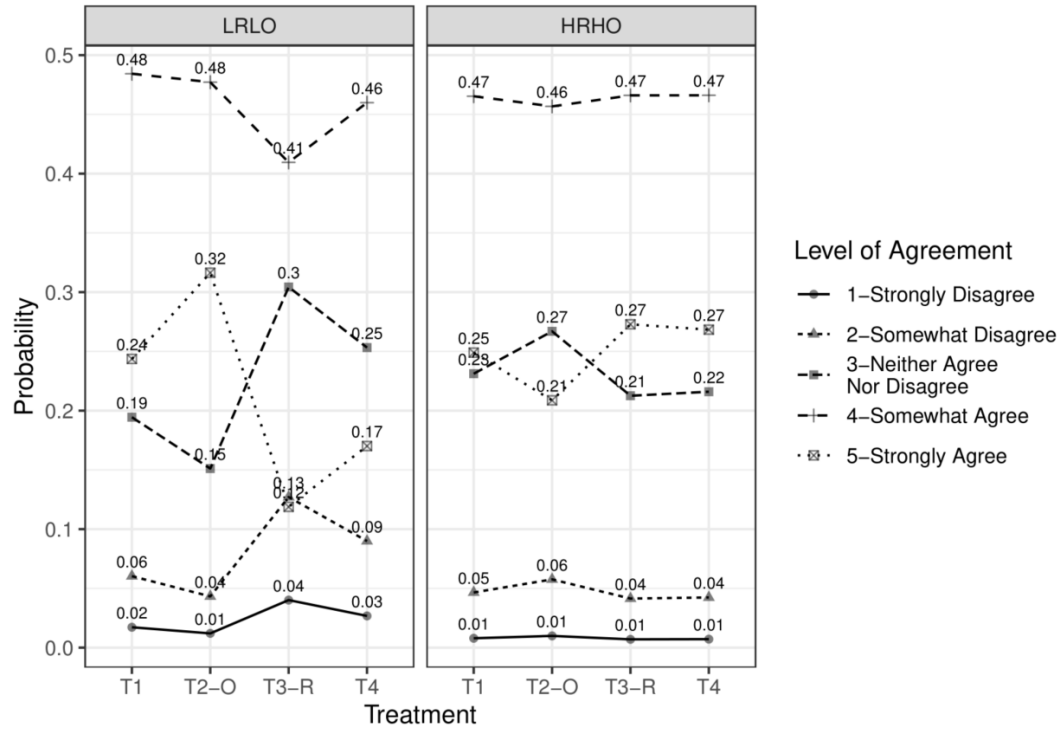


Figure 3. 3: Probability differences in perceived impact across treatments

3.6.3 Perceived Efficiency

Table 3.5 presents the ordinal logit estimates for the treatment effects on *perceived efficiency*. This measure appears to be more sensitive to both high and low values on both performance indicators of interest, although the influence of the overhead ratio prevails.

Compared to the no-treatment condition, reporting a low overhead tends to give individuals more and a high overhead less confidence in a nonprofit’s efficiency spending money, although the size of the coefficient is twice smaller for the HO-agencies. Information about charity ratings also affects the perceived efficiency: one’s awareness of a charity’s poor star-rating lowers their confidence in organizational efficiency, and the high rating increases the propensity to agree with the efficiency statement. Finally, introducing both a low rating and a low overhead

simultaneously works in the same direction as a low overhead alone, although yields a smaller-size coefficient, which, along with the statistically significant difference T4(RO) - T2(R), confirms the moderation effect of a low rating on the relationship between the *Overhead cost* and *Perceived efficiency*.

Table 3. 5: Perceived efficiency across treatments (ordinal logistic regression)

<i>Predictors</i>	LRLO		HRHO	
	<i>Log-Odds</i>	<i>Std. Error</i>	<i>Log-Odds</i>	<i>Std. Error</i>
T2 (O)	1.28 ***	0.18	-0.64 ***	0.18
T3 (R)	-0.75 ***	0.18	0.55 **	0.17
T4 (RO)	0.41 *	0.17	-0.26	0.17
1 2	-3.70 ***	0.25	-3.70 ***	0.23
2 3	-2.23 **	0.16	-2.00 ***	0.15
3 4	-0.04	0.12	-0.11	0.12
4 5	1.53 ***	0.14	1.58 ***	0.14
Same models with T2 (O) as the reference group:				
T1 (NRNO)	-1.28 ***	0.18	0.64 ***	0.18
T3 (R)	-2.03 ***	0.19	1.19 ***	0.18
T4 (RO)	-0.86 ***	0.18	0.38 *	0.18
1 2	-4.97 ***	0.26	-3.06 ***	0.23
2 3	-3.50 ***	0.18	-1.36 ***	0.14
3 4	-1.31 ***	0.14	0.53 ***	0.13
4 5	0.26 *	0.13	2.22 ***	0.15
Observations	873		873	
Cox & Snell's R ² / Nagelkerke's R ²	0.136 / 0.146		0.053 / 0.057	
Standard errors on parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$				

As in the previous case, in the reference condition, the initial probabilities describing the propensity to agree with the positive efficiency statement are similar for the two agencies with contrasting measured performance (Figure 3.4). Thus, the most likely responses are *Neither Agree Nor Disagree* (39.3% for LRLO and 35.4% for HRHO), *Somewhat Agree* (33.2% and 35.7%), and *Strongly Agree* (17.8% and 17.0%) for the two paired charities. Switching from T1 to T2, the probability ranking of response levels reverses for the LO organization to *Strongly Agree* (43.6%), followed by *Somewhat Agree* (35.2%), and then by *Neither Agree nor Disagree*

(18.3%). For the HO agencies, the probability ranking of the response levels remains almost the same except for *Somewhat Disagree* (↑) and *Strongly Agree* (↓) switching places. The probability to *Neither Agree nor Disagree* increased from 35.5% to 42.6% and the probability to *Somewhat Agree* dropped from 35.7% to 27.3%. In T3, the one-star rating weakened donor confidence in a nonprofit's efficiency relative to the no-information condition as the probability of declaring uncertainty raised from 39 to 49 percent and the probability of *Somewhat Agreeing* and *Strongly Agreeing* dropped from 33 to 24 and 18 to 9 percent respectively. A five-star rating, in turn, added some confidence as individuals ended up being nine percentage points more likely to *Strongly Agree* and nine percentage points less likely to be uncertain regarding an organization's efficiency. Finally, in T4, some improvements in the probabilities to *Strongly Agree* and *Somewhat Agree* with the efficiency statement can be observed for the LRLO-charity, even though they are smaller. For the HRHO, the probabilities become close to those in T1 as the effects of the rating and overhead offset each other.

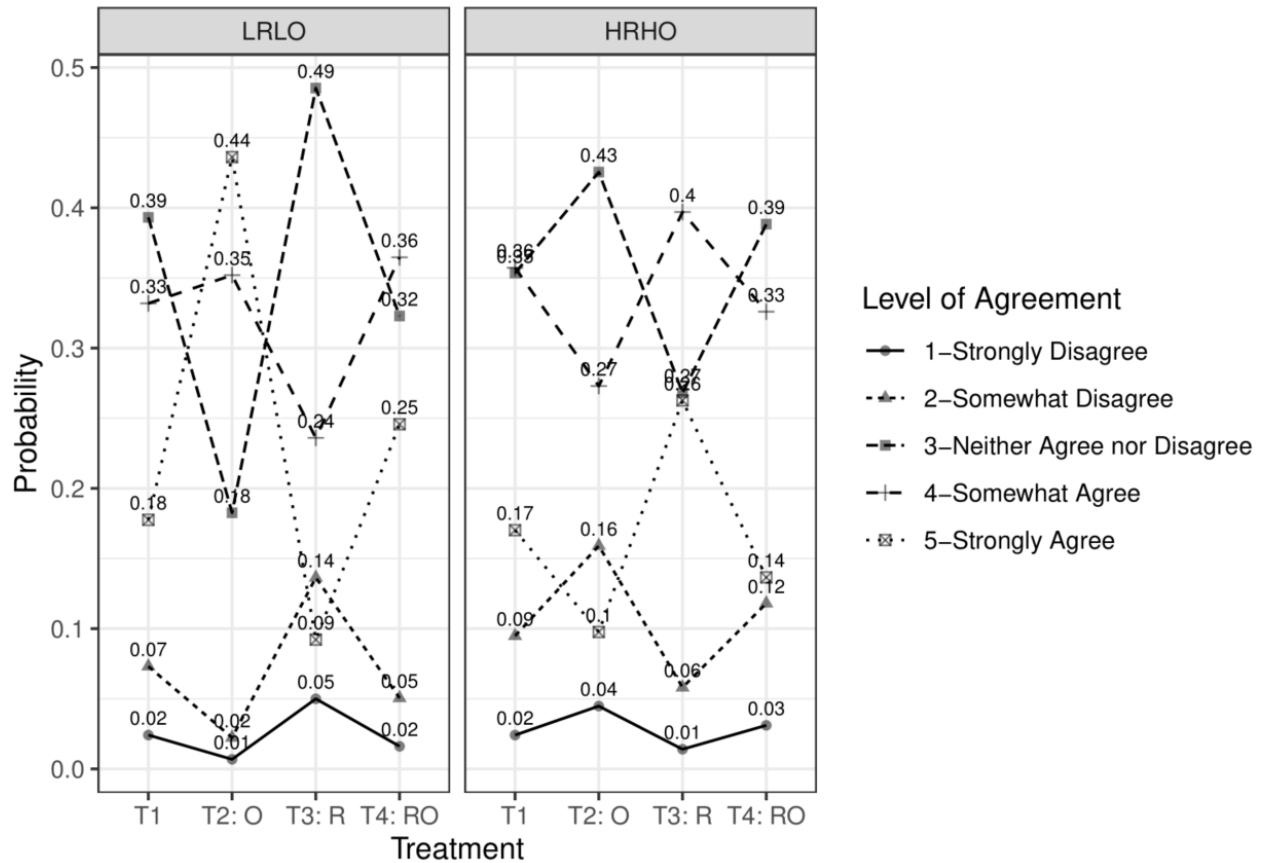


Figure 3. 4: Probability differences in perceived efficiency across treatments

3.6.4 Donor Trust

Figure 3.5 and table 3.6 show the differences across the treatments for *Donor Trust* in a nonprofit agency. As was the case with the perceived performance, learning about a low charity overhead statistically significantly increases donor trust in an agency relative to the reference condition. At the same time, a high overhead does not make a significant difference in trust. The estimates also show that having a five-star rating does not lead to a significantly different level of trust, whereas a one-star rating negatively affects donor trust. Finally, when both performance measures are presented in the report card, a low overhead ratio and a high rating offset each other's effects. The results partially confirm hypothesis H1b, H2b, and H3b.

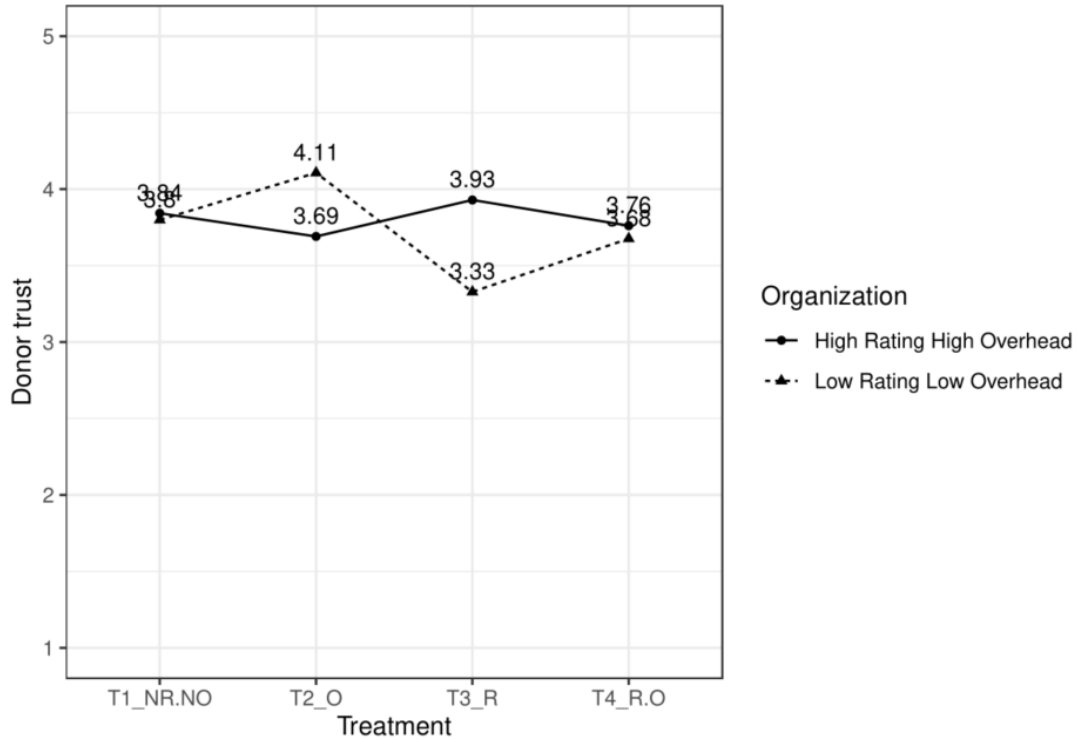


Figure 3. 5: Donor trust across treatment groups

Table 3. 6: Donor trust across treatment groups

	LRLO	HRHO
(Intercept)	3.80 *** (0.06)	3.84 *** (0.06)
T2 (O)	0.31 *** (0.08)	-0.15 (0.08)
T3 (R)	-0.47 *** (0.08)	0.09 (0.08)
T4 (RO)	-0.12 (0.08)	-0.08 (0.08)
Observations	873	873
R ² / adjusted R ²	0.096 / 0.093	0.011 / 0.008

Standard errors on parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

3.6.5 Donation Preference

As we can see in Figure 3.6, the reference group allocated the budget among the two agencies so that, on average, \$45 went to HRHO and \$55 to LRLO charities. Adding the performance measures to the report card leads to a statistically significant redistribution of donations among the charities (Table 3.7). Thus, facing the information on charity overhead

spending, experimental participants in T2 reallocated the estimated \$6.25 more to the LO organization widening the revenue gap between the two charities by the estimated \$12.5. The difference is significant at the five-percent level and confirms hypothesis H2c. The availability of the star-ratings on the report cards instead of the overhead measures yielded an even stronger effect with the estimated point difference of \$19.06 relative to the reference condition in favor of the highly-rated charity. The effect is highly significant and supports hypothesis H1c. Finally, reporting both measures again shifts donations to a highly rated charity, although its high overhead ratio attenuates the difference. The resulted difference is also highly significant, thus confirming hypothesis H3c.

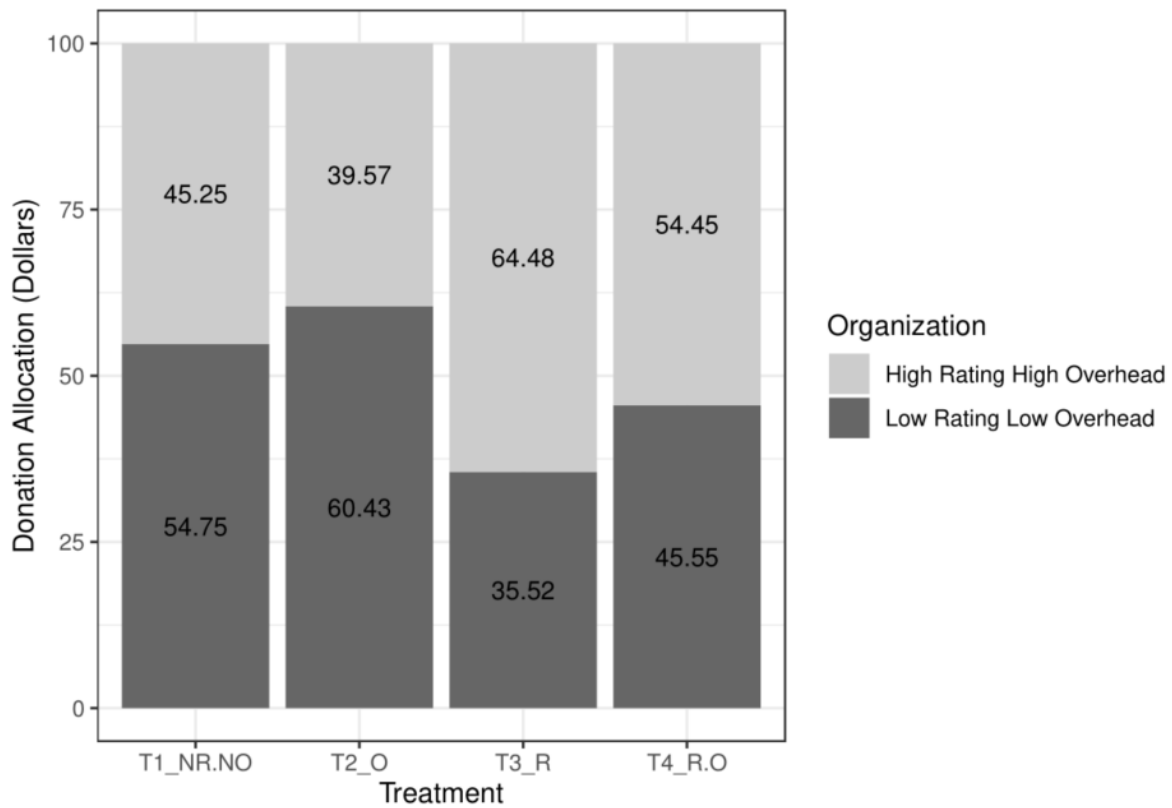


Figure 3. 6: Donor willingness to donate across treatment groups

Table 3. 7: Willingness to donate across treatments

	LRLO (T1 is ref. group)	HRHO (T1 is ref. group)	LRLO (T2 is ref. group)	LRLO (T3 is ref. group)
(Intercept)	55.34 *** (2.54)	44.66 *** (2.54)	61.60 *** (2.60)	36.28 *** (2.48)
Treatment T1: NRNO			-6.25 * (2.65)	19.06 *** (2.67)
Treatment T2: O	6.25 * (2.65)	-6.25 * (2.65)		25.32 *** (2.63)
Treatment T3: R	-19.06 *** (2.67)	19.06 *** (2.67)	-25.32 *** (2.63)	
Treatment T4: RO	-8.92 *** (2.63)	8.92 *** (2.63)	-15.18 *** (2.58)	10.14 *** (2.61)
Pair P2	3.44 (3.01)	-3.44 (3.01)	3.44 (3.01)	3.44 (3.01)
Pair P3	-1.38 (2.61)	1.38 (2.61)	-1.38 (2.61)	-1.38 (2.61)
Pair P4	-3.15 (2.55)	3.15 (2.55)	-3.15 (2.55)	-3.15 (2.55)
Observations	873	873	873	873
R ² / adjusted R ²	0.112 / 0.106	0.112 / 0.106	0.112 / 0.106	0.112 / 0.106

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

As presented in Tables 3.8-3.9, adding the measures of *perceived performance* and *trust* to the equation weakens the coefficients on all the treatment variables. In particular, controlling for either of the two variables renders the overhead condition to become statistically insignificant, suggesting that individual willingness to donate to a nonprofit is affected by its overhead spending entirely through perceptions of performance and trust, as the theory argues. A similar mediating effect is present on the path between the star-rating and giving behavior. However, even after accounting for perceived performance and trust, a highly significant direct effect remains.

Table 3. 8: Willingness to donate across treatments (Low-Rating-Low-Overhead)

	Model1	Model2	Model3	Model4
(Intercept)	55.34 *** (2.54)	50.91 *** (3.69)	11.42 * (5.05)	21.30 *** (4.89)
Treatment T2 (O)	6.25 * (2.65)	6.28 * (2.65)	1.87 (2.52)	3.26 (2.56)
Treatment T3 (R)	-19.06 *** (2.67)	-18.81 *** (2.66)	-14.07 *** (2.54)	-14.75 *** (2.59)
Treatment T4 (RO)	-8.92 *** (2.63)	-9.14 *** (2.63)	-8.59 *** (2.47)	-8.12 ** (2.53)
PairP2	3.44 (3.01)	2.23 (3.02)	2.90 (2.83)	2.15 (2.89)
PairP3	-1.38 (2.61)	-0.77 (2.62)	-0.32 (2.46)	-1.05 (2.51)
PairP4	-3.15 (2.55)	-3.73 (2.55)	-2.05 (2.40)	-3.41 (2.44)
Emotional utility		-1.19 (0.94)	-1.96 * (0.89)	-1.55 (0.91)
Familial utility		3.71 *** (1.10)	2.82 ** (1.04)	2.01 (1.07)
Familiarity		-0.65 (1.26)	-2.10 (1.19)	-1.63 (1.22)
Perceived performance			11.97 *** (1.11)	
Trust				9.46 *** (1.08)
Observations	873	873	873	873
R ² / adjusted R ²	0.112 / 0.106	0.124 / 0.115	0.228 / 0.219	0.196 / 0.187

Standard errors on parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 3. 9: Willingness to donate across treatments (High-Rating-High-Overhead)

	Model1	Model2	Model3	Model4
(Intercept)	44.66 *** (2.54)	32.29 *** (3.58)	2.87 (4.84)	8.14 (4.88)
Treatment T2 (O)	-6.25 * (2.65)	-5.82 * (2.62)	-3.27 (2.53)	-4.68 (2.55)
Treatment T3 (R)	19.06 *** (2.67)	19.28 *** (2.64)	17.55 *** (2.54)	18.49 *** (2.57)
Treatment T4 (RO)	8.92 *** (2.63)	9.62 *** (2.61)	10.12 *** (2.51)	10.08 *** (2.54)
Pair P2	-3.44 (3.01)	-4.03 (2.99)	-3.77 (2.87)	-3.84 (2.91)
Pair P3	1.38 (2.61)	1.12 (2.60)	0.97 (2.50)	0.69 (2.53)
Pair P4	3.15 (2.55)	2.97 (2.52)	1.71 (2.42)	2.56 (2.45)
Emotional utility		2.94 ** (0.95)	1.71 (0.92)	2.39 ** (0.92)
Familial utility		1.36 (1.10)	0.32 (1.06)	0.30 (1.08)
Familiarity		1.26 (1.36)	-0.61 (1.32)	0.33 (1.33)
Perceived performance			10.14 *** (1.18)	
Trust				7.75 *** (1.10)
Observations	873	873	873	873
R ² / adjusted R ²	0.112 / 0.106	0.139 / 0.130	0.207 / 0.198	0.186 / 0.177

Standard errors on parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 3.10 presents the results of testing the moderating effects of *mission valence*, *general trust in nonprofits*, and *altruism* on the relationship between the charity rating and willingness to give. First, the statistically significant at the five percent level coefficient on the interaction term for T3 and the indicator of similar causes (the interaction terms for T2 and T4 are marginally significant) confirm that mission valence influences the relationship between measured performance and giving decisions. As the table shows, when donors consider two nonprofit agencies addressing similar causes, the performance penalty (or reward) significantly shrinks in the conditions that involve the overall rating and might potentially increase in the overhead condition compared to the donation allocations among agencies with more disparate missions. However, the sign on the coefficient in treatment T3 delivers a finding that contradicts the hypothesized relationship as the highly rated charity gets a significant cut to its performance-based gain.

Second, the estimates also show that the moderating effect of *general trust* is significant in the rating only condition: as the level of individual trust in nonprofits in general increases, the

performance-based penalty/reward tends to shrink. This fully confirms hypothesis H5a that those who have relatively high levels of confidence in nonprofits tend to discount external performance rating information more heavily.

Finally, the moderating effect of *altruism* is also significant in the rating-only treatment group. However, the sign of the coefficient suggests that highly altruistic individuals tend to be less responsive to this performance measure than those with lower levels of altruism. This result is the opposite of the presented theoretical argument, so hypothesis 5b cannot be confirmed in its current formulation.

Table 3. 10: Willingness to donate across treatments

	Model 1 (Mission Valence)		Model 2 (General Trust)		Model 3 (Altruism)	
	LRLO	HRHO	LRLO	HRHO	LRLO	HRHO
(Intercept)	56.51*** (2.54)	43.49*** (2.54)	58.88*** (11.37)	41.12*** (11.37)	78.13*** (14.22)	21.87 (14.22)
Treatment T2(O)	2.01 (3.35)	-2.01 (3.35)	19.29 (15.12)	-19.29 (15.12)	4.12 (20.97)	-4.12 (20.97)
Treatment T3(R)	-23.93*** (3.55)	23.93*** (3.55)	-60.26*** (15.27)	60.26*** (15.27)	-61.44 ** (20.60)	61.44** (20.60)
Treatment T4(RO)	-13.27*** (3.44)	13.27*** (3.44)	-8.99 (15.23)	8.99 (15.23)	-30.92 (19.97)	30.92 (19.97)
Similar Cause T2(O) *	-3.97 (3.82)	3.97 (3.82)				
Similar Cause T3(R) *	10.19 (5.50)	-10.19 (5.50)				
Similar Cause T4(RO) *	10.74 * (5.37)	-10.74 * (5.37)				
General Trust T2(O) *			-1.05 (2.86)	1.05 (2.86)		
General Trust T3(R) *			-3.50 (3.81)	3.50 (3.81)		
General Trust T4(RO) *	9.98 (5.34)	-9.98 (5.34)				
General Trust Altruism						
General Trust T2(O)*Altruism					-5.67 (3.42)	5.67 (3.42)
General Trust T3(R)*Altruism					0.45 (5.00)	-0.45 (5.00)
General Trust T4(RO)*Altruism					10.27 * (4.97)	-10.27 * (4.97)
Observations	873	873	873	873	873	873
R ² / adjusted R ²	0.116 / 0.109	0.116 / 0.109	0.124 / 0.117	0.124 / 0.117	0.113 / 0.106	0.113 / 0.106

Standard errors on parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

3.7 Conclusions and Policy Implications

The growing prominence of nonprofit performance report cards has motivated a scholarly interest in estimating their efficacy. This study relies on a 4*2 mixed factorial design involving four between-group comparisons and two within-subject measures with a realistic decision setting to obtain answers to a few related questions. First, what effects do measures provided in charity ratings have on individual giving? Second, what perceptual determinants of charitable giving play a role in individual reactions to those measures? And third, how do individual characteristics affect the donor response to third-party nonprofit performance ratings?

As theorized, both composite star-ratings and efficiency ratios affect donor decision making. Contrasting values on either of the two measures can make a statistically significant difference in donor perceptions of those organizations' performance, levels of trust in them, and allocations of charitable contributions among them. However, "bad" ratings and "good" overhead ratios affect perceptions differently than "good" ratings and "bad" overheads. The study shows that donors are particularly sensitive to visibly low overhead spending ratios and the extremely poor composite rating while, at the same time, being not responsive to the excellent rating or relatively high overhead costs. Regardless of its questionable informational value, a low overhead ratio appears to effectively send a positive signal about organizational performance, including both effectiveness and efficiency, to individual donors. A high overhead ratio somewhat detracts from the perceived overall performance too, but its effect is smaller in magnitude and only significant in donor perceptions of organizational efficiency. Similarly, when considering the composite charity rating, a top star-grade might or might not improve donor perceptions of nonprofit performance in comparison to the no-information condition. A poor rating, however, damages the perceived performance. The patterns are similar for donor

trust in a nonprofit agency: a low overhead ratio increases trust, whereas knowing about a high overhead does not make individuals trust a nonprofit less. By contrast, the poor overall rating damages trust in a nonprofit while a high rating does not add to it in comparison to the no-information condition. Donor reactions suggest their high a priori expectations of nonprofit performance but low expectations of measured efficiency. As a result, individual donors demonstrate a willingness to punish a charity for poor overall rating by reallocating some of the charitable contributions to a more highly rated institution along with an inclination to reward a high measured efficiency. When the two indicators are simultaneously presented in the same report card, which is typically the case in practice, they interact and may send users contradicting signals. In donative decisions, however, the effect of the composite rating prevails.

Cumulatively, the findings suggest complex, nonlinear patterns in donor reactions to performance rating information and point to the importance of the content and design of nonprofit performance report cards. Not only do donors demonstrate asymmetrical uninformed expectations of nonprofit overhead spending and overall ratings, they tend to adjust their donative decisions according to those expectations and other individual characteristics such as general trust in nonprofits. Since quality rating agencies may play a role in shaping those expectations, it is important that they do it mindfully and cautiously, especially with respect to practical meaning and significance of different overhead spending ratios and differences in star-ratings.

The empirical confirmation of the interaction between the two performance measures is a disturbing finding with further practical implications since the overhead spending ratio is already incorporated in the rating. The fact that the effect of a poor overall performance rating can be

mitigated by advertising a low overhead ratio is worrying because it might push low-performers to engage in manipulating its reported overhead costs.

Such perceptions of the overhead ratio have three important ramifications for the nonprofit performance measurement. First, a better measure of organizational efficiency in the nonprofit sector is needed. Second, when incorporating the overhead ratio in a performance assessment models, its interaction with the overall performance grade must be accounted for in the design of nonprofit performance report cards so that it could be minimized if not eliminated. Third, this research indicates that individual donors may have a tendency to associate nonprofit efficiency with the overhead ratio while seeing the star-rating as a measure of its mission-related effectiveness. Since this is not entirely the case, the public either seems not to realize that a charity's performance rating already reflects its cost ratio or does not agree on the weight the measure has in the composite indicator, which is less plausible. Regardless of the reason, the phenomenon warrants more public education regarding the informational value of the overhead spending ratio as a performance indicator and its role in determining the overall rating.

3.8 Limitations and Future Research Directions

Although this study offers a systematic and focused inquiry into the individual-level effects of measures embedded in charity performance report cards and thereby advances our understanding of their regulatory potential, the paper is still only a first step to understanding the properties of this tool. The findings presented in this paper and its limitations point to new directions for research. First, by focusing on the extreme values of charity ratings and overhead ratios, this study has not explicitly modeled the relationship between the studied performance

indicators and outcomes. The experiment has been able to quantify some behavioral effects of the two studied performance measures, discover the likely nonlinear nature of the relationships, and confirm some interaction patterns. At the same time, further effort is required to reveal the form of the nonlinear patterns and all possible interactions in the whole range of the performance scales. Extending this work would allow us to make predictions outside of the scope of this analysis, for example, for nonprofits that score in the middle of the scale, and extend the scope of inquiry to understanding how such interactions as High-Rating-Low-Overhead or Low-Rating-High-Overhead further improve or damage individual perceptions and giving decisions. Second, the detected direct effect of performance ratings on donations bypassing perceived performance and trust points to new paths that charity composite performance ratings might operate through. Investigation of these new mechanisms requires additional theoretical inquiry and further testing, thus promising a more elaborate understanding of the ways publicized performance indicators influence the outcomes. Finally, further studies can focus on explaining the moderation role of mission valence, altruism, and other personal level characteristics. To this end, further examination of the mission valence can be improved through new research designs, including improved measurement of the concept, while investigating the role of altruism would benefit from additional theoretical work.

When considering the results of this study, it is also important to recognize the limitations that are inherent to the experimental design. Even though the experimental condition employed in this study was designed to strengthen its internal validity through approximating to a realistic decision-making situation, the experimental setting nonetheless remained artificial. In particular, the budget that experimental subjects allocated among competing charities was not their own money even though the decisions were consequential. Also, the subjects operated under a

constrained choice since the experimental design did not involve the option not to donate. Finally, the external validity of the study is limited by the characteristics of the panel provided by Qualtrics.

CHAPTER 4: PUBLIC CHARITY RESPONSE TO PERFORMANCE RATINGS

4.1. Introduction

There are different types and shapes of performance measurement systems (Barnow & Heinrich, 2010; Poister et al., 2014; Rowe, 2012). Barnow and Heinrich (2010) remind us that program evaluations, performance reports, benchmarking, report cards/consumer reports, and disclosure requirements are some of the existing approaches. Also, performance measurement in the public and/or nonprofit sector organizations, in the traditional sense, can be initiated and used by various actors/stakeholders and for multiple purposes. The spectrum of purposes is broad and includes responding to pressures for evidence of program effectiveness, improving communications, increasing public accountability, building public trust, recognizing good performance, making cross organizational comparisons, judging value created, supporting strategic planning, learning, allocating resources, and improving management and program outcomes (Behn, 2003). This list can be further continued, although most of the items are going to be only means to the one ultimate purpose, which is improving performance (Behn, 2003; Poister et al., 2014). In that regard, Poister et al. (2014) write that “expectations that performance management should contribute to higher levels of organizational performance and, in particular, better outcomes is almost universal among both proponents and critics of the performance movement” (p. 413).

Nonprofit performance report cards (third-party charity ratings) can be thought of as external performance monitoring systems because they are not initiated or implemented by the nonprofit (Gormley & Weimer, 1999). The purposes of report cards and performance measurement systems substantially overlap. One common goal is accountability to society and the contributing public for integrity, stewardship, and effective performance. Gormley (2003) writes that correcting information asymmetries and facilitating accountability between organizations and their various constituents are some of the key economic and political purposes of the report cards. Accordingly, the main rationale/driving force behind emergence of third-party charity raters stems from the need to improve public accountability, protect donors' interests, and guide informed donor decision making. For instance, Charity Navigator's mission statement explicitly emphasizes that the agency "works to guide intelligent giving" (Charity Navigator, 2015); Charity Watch's raison d'etre is in "providing donors with the information they need to make more informed giving decisions" (Charity Watch, 2020), and the BBB Wise Giving Alliance works to help "donors make informed giving decisions" (The BBB Wise Giving Alliance, 2015). The three major third-party charity evaluators as well as the new players in the field consistently claim their role is to increase allocative efficiency in the nonprofit sector.

Improving organizational performance is another key purpose that organizational report cards share with other performance measurement systems (Gormley & Weimer, 1999). Even though the missions of the charity raters make less explicit emphasis on improving performance of rated organizations themselves, Gormley (2003) explains that, ideally, organizations should "pay attention to report cards and adjust their behavior, in an effort to compete more effectively with other organizations that produce the same services" (p. 4). Evidence from the literature on the behavior of business firms, hospitals, and graduate schools supports this argument. Chatterji

& Toffel (2010) wrote that external ratings “beyond their stated objective of influencing investors, also influence the rated firms” (p. 918) helping reduce toxic emissions. Longo et al. (1997) concluded that “[p]ublic release of consumer reports may be useful not only in assisting consumers to make informed health care choices, but also in facilitating improvement in the quality of hospital services offered and care provided” (p. 1579) and described the observed improvements as “an important by-product” (p. 1579) of consumer reports. Gormley & Weimer (1999) argued that organizations’ behavior is the ultimate target of report cards to which organizational leaders “will attempt to respond in ways that advance the interests of their organizations” (p. 123). Nonetheless, scholars have “only begun to theorize how independent company ratings affect the organizations being rated, and have offered little guidance on how differences in firm characteristics influence response” (Chatterji & Toffel, 2010, p. 918) (p. 918). The evidence on what difference such information tools make is virtually nonexistent in relation to public charities that produce public goods.

Although report cards have potential to improve organizational performance, targeted organizations do not always respond as expected (Gormley & Weimer, 1999). In addition to self-improvement, reactions to report card may include nonresponse or a range of dysfunctional responses. Learning whether nonprofit report cards improve performance in public charities would fill a gap in the literature. Therefore, this research focuses on the following questions:

- 1) Do public charities change behavior in response to external performance charity ratings?
- 2) How do public charities respond to charity ratings?
- 3) What factors influence how charities respond to public ratings?

This study hypothesizes that public charities do pay attention to charity ratings and change in response to information that is released by raters. Externally provided performance

standards should facilitate organizational learning and stimulate performance adjusting behavior. However, it also anticipates that the responses are not uniform across public charities and may depend on managerial, organizational, and environmental characteristics.

The remainder of the paper proceeds as follows. The next section overviews the research on organizational responses to report cards in various fields and industries. After that, an overview of the theoretical frameworks that explain the mechanics behind the relationship between ratings and organizational change is provided. The paper continues by applying the theory to charitable organizations, presenting the models explicating how public charities adjust to charity ratings, and describing how the proposed theory can be tested empirically. Then it presents the results of empirical analysis, conclusions, and limitations.

4.2 Literature Review

4.2.1 Do Organizations Change Behavior in Response to Ratings? Evidence from Education, Healthcare, and the Corporate Sector

Evidence on how independent performance ratings affect organizational behavior emerged in the early 1990s. Many studies of corporate environmental ratings, hospital ratings, school/university ratings, corporate social responsibility ratings, corporate/municipal credit ratings evidence supports the claims of performance measurement theory that external performance monitoring systems affect organizational behavior and performance, although not always as intended.

Much early research focused on educational institutions. Several studies on public schools in North Carolina, Florida, Texas, and Kentucky report that schools adjust to their public

ratings by improving performance (student test scores, pass rates, and various subject-specific skills) (Gormley, 2003). A qualitative study of how eight top business schools reacted to Business Week magazine's rankings of U.S. business schools provides a detailed account of how organizational members use cognitive tactics to cope with identity-threats created by unfavorable ratings. In particular, members selectively focused their attention on favorable aspects of their organizations' identities to restore positive perceptions about their organizations and reinterpreted rankings as misleading representations (Elsbach & Kramer, 1996). A quantitative study of business schools in the context of the U.S. News and World Report rankings reported that schools responded to rankings through organizational change (Martins, 2005). The variation in change depended on the discrepancy between rankings and managers' own beliefs about their schools' standing as well as managers' perceptions of the impact of the rankings. Similarly, Espeland and Sauder (2007) also found evidence of behavioral adaptation in law schools in response to being evaluated by the U.S. News and World Report rankings.

The hospital industry has also showed making performance adjustments in response to public rating information. A series of articles reported that the introduction of rankings or ratings intended to increase consumer-patient awareness improved hospital policies, procedures, and outcomes, including declines in surgery mortalities in New York hospitals. Peterson, DeLong, Jollis, Muhlbaier, and Mark (1998) found that surgery outcomes improved significantly: mortality rates declined faster than the national average in New York after the New York State Department of Health started to publicly release scorecards/mortality reports, and "NY had the lowest risk-adjusted bypass mortality rate of any state in 1992" (p.993). Similarly, Longo et al. (1997) found that following the publication of a consumer report, hospitals adopted policy changes and implemented improvements, "especially in competitive markets and areas of

care identified as possibly ‘out of alignment’ with care provided by high-quality performing peers” (p.1582).

A substantial body of literature examines market, social, or environmental performance of various for-profit firms. Graham (2000) describes several companies making rapid changes in products and completely legal practices in response to health and safety information in order to avoid public humiliation, even when they denied the rationale behind the disclosure requirements. Another piece of evidence comes from the restaurant market. When Los Angeles County required restaurants to publicly display grade cards of their hygiene inspections in the format of standard grade cards, Jin and Leslie (2003) found that restaurants responded with service quality improvement to avoid revenue loss. Firms also respond to corporate environmental ratings. Firms that initially scored poorly, improved more than firms that were not rated or initially rated higher (Chatterji & Toffel, 2010). Sharkey and Bromley (2014) found that even unrated firms improved in response to environmental performance ratings, which are capable of driving “field-wide change when only some firms are formally subject to evaluation” (p.64).

In sum, organizations that produce privately consumed goods and services respond to being monitored and rated and those responses vary across organizations in terms of how they respond and how much change they demonstrate. At the same time, the nonprofit literature is virtually silent with respect to organizations’ sensitivity to external performance monitoring and charity ratings. Sometimes, nonprofits advertise their high ratings in their communications with the public. For example, one charity writes on its website that it “strives to earn the highest charity ratings to give you assurance that your support will be used effectively and efficiently” (Environmental Defense Fund, 2020). Another charity’s message to its constituents states that its

“work earns wide recognition from independent charity evaluation agencies, including a 100% fundraising efficiency rating from Forbes, a spot on Charity Navigator’s list of the “10 Best Charities Everyone’s Heard Of” (Direct Relief, 2020). Nonetheless, we do not know whether such responses generalize to the whole population of rated charities. Therefore, research addressing charity response to performance report cards would greatly inform nonprofit sector theory and practice.

The following sections explicate the theory of the relationship between external performance monitoring/rating and organizational behavior and applies the outlined theoretical statements to a subset of the nonprofit sector - public charities that produce public goods/services and are funded through voluntary public contributions.

4.2.2 Theory of Organizational Response to External Performance Monitoring

Economic theories of organizational behavior focus on ideas of information asymmetry, bounded rationality, organizational slack, agency problems, information search cost, attention focus, and customer response. The “lemons” framework (Akerlof, 1970) explains how information asymmetry regarding product quality (when sellers have more information about their product than buyers) drives dishonesty and market inefficiency. The theory shows that dishonesty on behalf of sellers and the corresponding uncertainty of buyers “tend to drive honest dealings out of the market” (p. 495). In the markets with information asymmetry, the cost of dishonesty would “include the loss incurred from driving legitimate business out of existence.” (495). In the nonprofit sector, which is characterized by a high degree of information asymmetry

between those who produce goods and services and those who pay for them, this would also lead to deterioration in performance if cheating generated a surplus to those in control.

Viewing the donor-nonprofit relationship through the prism of principal-agent theory yields similar predictions (Moe, 1984). The principal (donor) would expect the agent (charity) to produce outcomes that satisfy the principal's objectives. However, "there is no guarantee that the agent, once hired, will in fact choose to pursue the principal's best interests or to do so efficiently" (Moe, 1984). Information asymmetry between the two creates moral hazard for the agent to pursue their own agenda, which leads to a conflict of interest.

By removing the residuals that could be distributed to the owners from the structure of a nonprofit organization, the nondistribution constraint is expected to counterbalance the incentive to engage in dishonest and compromise on quality. But this does not solve the performance problem due to existence of organizational slack. Cyert and March (1963) define organizational slack as "payments to members of the coalition¹² in excess of what is required to maintain the organization" (Cyert & March, 1963, p. 36), or, in other words, the difference between the actual spending and "the true minimum cost of service provision" (Moe, 1984, p. 763). According to Hirschman (1970), slack can also be viewed as "a gap of a given magnitude between actual and potential performance of individuals" (p.14).

Slack can exist in many forms. Unabsorbed slack can accumulate in uncommitted liquid resources, while absorbed slack can reflect excessive costs, such as production inefficiencies, policies, wages, services, and personal perquisites (Cyert & March, 1963; Singh, 1986). All the forms of it are documented in nonprofits (Kelly, 1998). Cyert and March (1963) argues that slack is "useful in dealing with the adjustment of firms to gross shifts in the external environment"

¹² Coalition may include managers, workers, other paid functionaries, suppliers, customers, lawyers, tax collectors, regulatory agencies, volunteers, donors, donees, etc. (Cyert & March, 1963, p. 27)

(p.37). In a favorable environment, a well-performing organization accumulates slack. However, when it faces adversity and potential failure, organizational slack provides a cushion that helps the organization to adapt to the shift in the environment and survive.

To mitigate the effect of information asymmetry and restore optimality, the customer (principle) faces the challenge of how to identify quality (or reveal the agent's privately held information). Therefore, the less informed party could employ performance monitoring. Information about an unsatisfactory quality of an organization's output, could lead to customer reaction in the form of "exit" (causing a loss of revenue), or "voice" (through expressing complains) (Hirschman, 1970). Both reactions, as well as an emergence of external monitoring can become threatening exogenous events that would initiate organizational response to unfavorable conditions through attention focus mechanism, information search, upward adjustment of aspirations, and absorbing slack resources (Cyert & March, 1963; Singh, 1986).

The empirical literature on the organizational effects of report cards supports the role of economic incentives, social/political pressure, and attention focus in determining organizational motivations to improve their performance in response to external performance monitoring. One strand examines information asymmetry, information search cost, reputation, embarrassment and shame as main mechanisms that stimulate organizational change in response to external assessments. Thus, Gormley (2003) argues that report cards influence the behavior of organizations and lead to service delivery improvements because they "shape the choices that consumers or purchasers make, resulting in a shift of organizational market shares" (p.13) and because public information on poor performance causes embarrassment in evaluated organizations. Using a number of examples from government mandatory disclosure regulations, Graham (2000) argues that release of negative, shaming information/ratings to consumers makes

companies change their products even if they disagree with such information: “The company's reputation, hard to build and easy to destroy, is at stake” (p.37). Jin and Leslie (2003) present empirical evidence that economic incentives stimulate restaurants with poor hygiene to improve after hygiene grade cards are mandatorily disclosed to consumers. This effect is expected through reducing search costs to consumers, mitigating information asymmetry, and altering the nature of competition among restaurants. The researchers confirmed their arguments by presenting empirical evidence that restaurants indeed responded to the introduction of hygiene grade cards with hygiene quality improvements and, therefore, a correspondent average increase in inspection scores. Jin and Leslie (2009) show that restaurant hygiene grade cards can facilitate consumer learning about a firm’s unobservable characteristics (e.g. a restaurant’s hygiene quality) and its reputation formation process. Because increased reputation could be instrumental in generating resources for the firm, whereas a loss of reputation associated with poor performance could entail long-term costs (Lewis, 2014). In sum, the economic perspective predicts that external performance assessment, in the form of either embarrassing information or recognition of excellence, would cause subsequent performance improvement.

Several behavioral models, drawing on organizational/social identity, performance feedback, behavioral, stakeholder, and institutional theories help understand how organizational perceptual mechanisms, information processing limitations, and environmental pressures shape organizational focus and reactions. Thus, two studies of the effects of rankings on behavior of US graduate schools of business, one qualitative and one quantitative, employ organizational/social identity perspective that focuses on microprocesses of organizational adaptation. Drawing on social identity, self-affirmation, and impression management theories, Elsbach & Kramer (1996) explain organizational response to rankings by treating them as

“events that threaten their perceptions of their organization’s identity” (p. 442). In the absence of external rankings, schools’ members shape their own self-image that allows them to promote their own ideas about their organizations’ important identity attributes and relative standing in the industry. Business Week rankings emerged as powerful external institution that imposed evaluation of business schools against objective and uniform criteria, which challenged “the merit or importance of core distinctive and enduring organizational traits” (p. 444) and “dramatically disrupted the status quo that these schools had long enjoyed, creating an organizational identity threat” (p. 444). The organizational identity management framework suggests that such disruption and emergence of a threat is followed by members’ efforts to restore and protect positive perceptions of their organizational and social identity, which might range from ignoring or resisting rankings to using cognitive tactics to maintain positive sense of self. The latter is done through reinterpreting their standing relative to rankings using selective categorizations (strategies), favorable comparisons, or positive highlights of identity traits not captured by rankings. Although this analysis doesn’t uncover any measurable responses along quantitative metrics and focuses on perceptual, cognitive, and psychological ways to cope with external institutional pressures, it shows that organizational members “care about how their organizations are described and also how they compare with other organizations” (p. 468) and protect their personal and their organizations’ social identities. In addition, the discovered sensemaking activities help organizations focus attention on what they should be doing and why, thus pointing out the importance of a constructive change process and symbolic management. Elsbach & Kramer (1996) write that “using selective categorization processes creatively can help organizations decide not only where emerging opportunities lie, but also what the appropriate and useful responses to them are” (p. 474).

Drawing on behavioral and performance feedback theories, Lewis (2014) argued that managers often lack information to make optimal choices and need coping mechanisms to deal with uncertainty about the future. A third-party performance rating sets fixed standards for performance, which can reduce uncertainty. In the absence of complete information necessary to make a rational decision, a boundedly rational organization may adopt the rating as a decision rule. They write “following a performance standard established by a rating may be superior to alternative decision rules as it does not require firms to revisit the decision each year and thus reduces the costs of information search and cognitive processing” (p.11). Unfortunately, this theory also implies that an external performance benchmark can also cause a highly performing organization to lower its performance. As Lewis (2014) argues, “just meeting the benchmark may in fact be the optimal response” (p.11), so a positive recognition can decrease a firm’s performance aspirations and its further performance at least to the satisfactory level determined by the external benchmark.

Another useful theoretical lens that deepens our understanding of organizational reactions to external performance monitoring and provides additional arguments to expect organizational adaptation to ratings is presented by Espeland and Sauder (2007) and draws on the methodological concept of reactivity. Known since at least the 1920s as the Hawthorne or observer effect, reactivity suggests that “individuals alter their behavior in reaction to being evaluated, observed, or measured” (Espeland & Sauder, 2007, p. 6). Reactivity is a well-known methodological concern in the social sciences, but Espeland and Sauder investigated the phenomenon in substantive terms by analyzing the reactivity of law schools to the U.S. News and World Report rankings. Taking a case-study approach to studying the consequences of reactivity for organizational behavior in the presence of external rankings, they discovered such

organizational reactions as “redistribution of resources, redefinition of work, and proliferation of gaming strategies” (p.3). To explain the mechanisms of organizational reactivity to rankings, they used the notions of self-fulfilling prophecy and commensuration. Similarly to the earlier idea about “how economic theory shapes the economy” (p.6), rankings too “change how people make sense of situations” (p. 10). Specifically, they define self-fulfilling prophecies as “processes by which reactions to social measures confirm the expectations or predictions that are embedded in measures or which increase the validity of the measure by encouraging behavior that conforms to it” (p.11). According to the scholars, rankings create certain expectations about schools and those expectations amplify their effects.

Whereas self-fulfilling prophecies affect behavior through altering expectations, commensuration alters individual cognition. Espeland and Sauder (2007) argue that “commensuration shapes what we pay attention to, which things are connected to other things, and how we express sameness and difference” (p. 16). Commensuration effects shape individual attention through cognitive mechanisms of simplifying information and unifying and distinguishing the targeted objects by constructing shared metrical relationships. Therefore, rankings “challenge ... fragmentation by reducing distinctiveness to magnitude” (p.19) that makes it “much harder to make status claims not supported by rankings or to sustain identities that are not linked to rankings” (p.19). Finally, presence of rankings encourages people to reflect on the ontology and relationship between the numbers and what they measure. One stance of the “reality” often adopted by those who know little about the methodologies underlying rankings is that “the social relationship that is measured is as real as a physical object” (p.21). The scholars write that “most are uninterested in ranking methodology and simply assume that rankings measure something real about the schools” (p. 21).

Overall, by developing and empirically verifying this constructivist view of rankings, Espeland & Sauder (2007) show that such monitoring systems can create powerful behavioral effects in organization by merely altering and framing the views of the reality for targeted audiences in a certain way. It is also useful for studying external performance monitoring and public disclosure to note the results of adaptation to rankings discovered by the researchers. As they show, the studied schools responded with efforts to maximize rankings through budgetary reallocations, redefinition of policies and procedures, and manipulation strategies. At the same time, such responses may be dysfunctional stimulating achievement of formal improvements only on the metrics used to construct rankings. Performance measurement scholars have long noted that poorly designed performance measurement systems can encourage undesirable behaviors (Poister et al., 2014). The performance measurement literature is rich in examples of dysfunctional responses, including nonparticipation, goal displacement, gaming, number manipulating, outright cheating, or challenging the validity and usefulness of the performance measures/system (Gormley, 2003; Gormley & Weimer, 1999; Poister et al., 2014). Organizations under pressure might engage in symbolic responses (Sauder & Espeland, 2009), or such goal displacement activities as “teaching to the test” type of behavior when “students may know more facts, while their ability to interpret the facts suffers” (Gormley, 2003, p. 14). They might game the system or even get involved in outright cheating. Either way, the theory of reactivity predicts improvement on the metrics that affect ratings, although not necessarily beyond that.

Finally, a consistent and overlapping with the discussed above theories approach to understanding the power of performance report cards is through the institutional perspective. The reviewed work has already recognized ratings and rankings as powerful institutions capable of stimulating organizational change. Chatterji & Toffel (2010) admitted the importance of

institutional expectations for legitimacy and survival of an organization. Sharkey & Bromley (2015) explained indirect effects and diffusion through institutionally altered processes of social construction. Martins (2005) emphasized “theoretical connection between cognition and institutional research” (p. 704), recognizing “rankings as important sources of institutional isomorphic pressures” (701). The author blamed the institutionalization of rankings as a possible reason of his nonfinding that managers’ perceptions of rankings’ validity was not a significant determinant of organizational change. He wrote that “the rankings have become institutionalized in this organizational field, rendering managerial assessments of the rankings secondary to institutional pressures from the rankings to conform” (p.714). Institutions can impose intense pressures and expectations, thus threatening organizational survival and becoming constraining forces that modify organizational characteristics in the direction of environmentally determined homogenization, which DiMaggio and Powell (1983) call institutional isomorphism.

DiMaggio and Powell (1983) describe three mechanisms driving institutional isomorphism: coercive, mimetic, and normative. Coercive isomorphism is associated with “both formal and informal pressures exerted on organizations by other organizations” (p.150), and can be pushed by a common policy environment, centralization or coordination processes in an organizational field, rituals of conformity to wider institutions, or even persuasion. Mimetic processes work through imitation and modeling on other organizations in response to uncertain environments. And, finally, the normative pressures are created by professionalization, suggesting that “organizational fields that include a large professionally trained labor force will be driven primarily by status competition” (p.154).

Further, the institutional view suggests an idea of “institutional duality” (Hunter & Bansal, 2007), claiming that organizational formal and informal structures are often “loosely

coupled” (Sauder & Espeland, 2009). As a result, organizations develop policies to improve attributes captured by external assessments but may also engage in manipulating statistics, redefining goals, or innovating with gaming techniques. They write that “To secure legitimacy and conform to general expectations, organizations may develop symbolic responses to environmental pressures without disrupting core technical activities” (Sauder & Espeland, 2009, p. 63). Manipulation strategies in a given field diffuse quickly as organizations are attentive to what others do to improve their standing in ratings. Only a few organizations that have little to lose and limited opportunities to improve may ignore publicized assessments and accept their inferior performance status, “reinterpreting the stigma of rankings as an honorable sacrifice” (p.78).

In summary, the outlined theoretical account of the mechanisms through which external performance monitoring systems influence organizational behavior shows how institutionalized, objective, and shared metrics permeate boundaries between organizations, become internalized by organizations, and pressure organizations toward change. They explain how resource dependence, competition, uncertainty, sensemaking and the fact of being evaluated motivate organizations to improve their attributes or resist. These theoretical statements offer arguments for developing a theoretical framework describing adaptation of public charities to charity ratings as discussed in the following section.

4.2.3 Public Charity Response to Third-Party Performance Ratings: Theory and Hypotheses

The logic behind these theoretical mechanisms applies to the behavior of public charities facing third-party external evaluations, such as report cards/ratings are publicized, and informs a

framework of a charitable organization's response to third-party ratings. Nonprofit charitable organizations may not accept the idea of ratings, may prefer different bases for performance evaluation, or have strategic priorities divergent from the dimensions emphasized by ratings, but ratings could still change their performance.

Most of the reviewed literature admits the fundamental role of economic incentives in facilitating performance improvements with publicized rating information, even when other, less visible mechanisms may also be at work (Chatterji & Toffel, 2010; Gormley, 2003; Graham, 2000; Jin & Leslie, 2003; Longo et al., 1997). As in markets of private goods (e.g. restaurants, healthcare, or education), information asymmetry is present in the relationship between nonprofits that produce public goods and their key funding stakeholders, which prevents funders “from knowing when to believe suppliers’ claims about product attributes that are not directly observable” (Chatterji & Toffel, 2010, p. 917). As third party assessments are increasingly gaining public attention and valued by their audiences (Longo et al., 1997; Sauder & Espeland, 2009), they can mitigate the depth of the information asymmetry problem and alter the nature of competition for resources (contributions) among charitable organizations. Third-party ratings provide easy-to-access evaluations that summarize performance using easy-to-comprehend aggregate measures and enable quick and simple comparisons across charities of interest. Thereby, third-party charity ratings reduce search costs and costs of learning for stakeholders and may influence funders’ decisions to contribute to some charities more than to others; in other words, introduction of independent ratings can influence organizational market shares (Gormley, 2003; Jin & Leslie, 2003). Thus, to maintain or increase their market shares, organizations may try to improve their performance (Jin & Leslie, 2003; Longo et al., 1997). Publication of charity

performance ratings, therefore, can emerge as threatening event and influence charities to adjust/improve their subsequent performance on the publicized metrics.

Public ratings can affect an organization's reputation (Jin & Leslie, 2003), so regulation by shaming (Gormley, 2003; Graham, 2000) also applies in the nonprofit sector where trust goods and services are produced. Even without changes in market shares, a charity with a poor external assessment may face embarrassment and public humiliation with potential consequences for its reputation. Nonprofit managers will want to improve their organizations' standing with their external evaluators. Other theories suggest similar basic expectations. Stakeholder theory dictates that "the identity of stakeholders and the nature of their requests influence firm responsiveness" (Chatterji & Toffel, 2010, p. 918). Nonprofits that depend on donations will want good external ratings, especially if those ratings influence giving. Boundedly rational nonprofit managers will not be able to determine the economically optimal level of performance, so they may redefine optimality in terms of charity ratings and use them as fixed performance standards that reduce uncertainty and require a lower cognitive effort. Finally, institutional and measurement reactivity theories also predict that publicizing of charity ratings will have significant effects on the charity performance scores. Hence, there are several compelling reasons to put forward the following hypothesis:

H1: Nonprofits that receive a third-party rating will subsequently improve their measured performance¹³

¹³ Performance, as a multidimensional concept, is represented by several measures, as discussed in the Measures section. Therefore, each hypothesis breaks down into several sub-hypotheses – one for each operational indicator of performance

Variations in Charity Response to Third-Party Ratings

Organizations vary in how they respond to external performance monitoring.

Organizational, institutional, and environmental factors may condition the amount of change in performance that ratings elicit. For example, Chatterji & Toffel (2010) showed that regulatory stringency and organizational efficiency moderated the influence of corporate environmental ratings on firms' subsequent environmental performance. Lewis (2014) found that normative pressures from local communities, industry-specific risk profiles, and prior financial performance conditioned the relationship between corporate social responsibility ratings and improvements in corporate social performance. These and other theories suggest a few contingencies relevant to the behavior of public charities under third-party performance monitoring. They suggest how a public charity's response to external performance ratings should vary under various organizational or environmental conditions.

First, economic theory suggests that performance improvement is costly and should be justified by expected benefits. Poorly measured performance of an organization may reflect a relatively large amount of slack resources, on which to rely for improvements, compared to an organization with high performance grades. Chatterji & Toffel (2010) argue that organizations with different levels of performance face different sets of opportunities for improvement: poorly performing/rated organizations "face lower marginal costs of improving their performance" (p. 922) and are more likely to implement lower cost but higher impact improvements than their more highly performing peers. The higher the initial performance is, the harder and costlier it is to further improve and the smaller the increments are. Poorly rated firms face lower cost improvement opportunities and greater potential benefits than their more highly rated peers and therefore are more likely to improve or show greater levels of organizational change (Chatterji &

Toffel, 2010; Lewis, 2014; Martins, 2005). Thus, charities with poor initial ratings will “have a greater opportunity to exploit low-hanging fruit” (Chatterji & Toffel, 2010, p. 922) and improve performance more than those with higher initial ratings. Hence it is reasonable to hypothesize that:

H2: *Charities initially rated poor will demonstrate higher levels of improvement than those initially rated higher (except those initially rated excellent)¹⁴*

The second hypothesis is consistent with the behavioral and performance feedback theory, although taking its assumptions fully into account adds an additional contingency. Following the logic explicated in Lewis (2014), boundedly rational organizational managers will incorporate external ratings as decision rules that will help optimize performance. Such a decision rule will create stimuli for a charity to just meet a performance benchmark set by the rating institution. Hence, charities with low ratings will attempt to improve their performance indicators to the norm set by the rating agency. On the other hand, given that just meeting the standard is construed as the optimal performance level, an organization initially scoring above the mark is likely to somewhat reduce its subsequent performance through absorbing part of the resources as slack:

H3: *Charities that initially receive the highest rating will subsequently reduce their performance on measured indicators*

Confirming the logic advanced by economic theory, prior research emphasizes the role of market competition as a significant factor influencing organizational responsiveness to public ratings. For instance Jin and Leslie (2003) theorized that hygiene ratings cause improvements in restaurant quality through the competition mechanism. The changes in quality of hospital care

¹⁴ The rating scales are outlined in the Methods section

found by Longo et al. (1997) were found to be especially pronounced in competitive markets. Given that nonprofits also compete for scarce charitable contributions, the nature of competition in a peer group (subfield) may determine how sensitive their behavior is to third party charity ratings. Specifically, the expectation is to see more performance improvement in response to ratings in more “crowded” fields of charitable activity – where competition for resources is fiercer. This expectation translates in the following hypotheses:

H4.1: *Nonprofits in fields with more competition will improve more in response to ratings than organizations in markets with less competition.*

H4.2: *Charities that rely more on public contributions will improve more than organizations that rely on contributions to a lesser extent.*

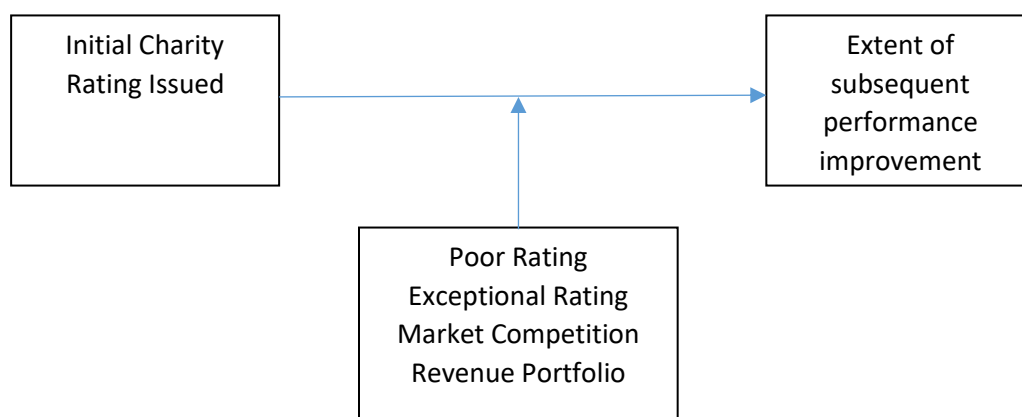


Figure 4. 1: Research model

4.3 Data

To test the hypotheses about how third-party ratings influence monitored performance of public charities, this research uses a simple random sample of report cards for the rated charities

from Charity Navigator. The sample contains 10345 observation for 841 charities between FY 2000-2018.

The choice of the rater for this analysis is strictly instrumental. Charity Navigator was founded in 2001 and, despite being one of the youngest third-party performance raters in the field, it still has a substantial history of producing ratings, which satisfies the data needs for this analysis. More importantly, it is the largest rater at this time grading over 9,000 public charities compared to 1,300 organization evaluated by the BBB Wise Giving Alliance, and about 575 nonprofits rated by Charity Watch. Additional factors that determined the choice of the rating context was availability of the historical rating data.

Unlike Charity Watch, which provides access to most of its ratings to its paid members only, Charity Navigator ratings are open access. Thus, the CN report cards are cheaper and easier sources of information to charities' constituents. Further, by funding its operations through public contributions and providing zero-cost access (Charity Navigator, 2013), Charity Navigator is preferable to BBB Wise Giving Alliance, which receives 82 percent of its income from Charity Seal license fees (BBB Wise Giving Alliance, 2013). Choosing a rating agency whose selection and evaluation methods are independent from motivations of rated organizations is important for minimizing potential self-selection issues.

Finally, besides a simple four-star rating, the Charity Navigator uses a convenient for quantitative analysis interval level scale, as presented in Table 1.

Table 4. 1: Charity Navigator’s grading scales

Numeric Score	No. of Stars	Qualitative Rating	Description
91-100	★★★★	Exceptional	Exceeds industry standards and outperforms most charities in its Cause.
80-90	★★★★☆	Good	Exceeds or meets industry standards and performs as well as or better than most charities in its Cause.
70-80	★★★☆☆	Needs Improvement	Meets or nearly meets industry standards but underperforms most charities in its Cause.
55-70	★☆☆☆☆	Poor	Fails to meet industry standards and performs well below most charities in its Cause.
< 55	0-Stars	Exceptionally Poor	Performs far below industry standards and below nearly all charities in its Cause.
	Donor Advisory	No Rating	Serious concerns have been raised about this charity which prevents the issuance of a star rating

Charity Navigator issues three types of ratings: accountability and transparency ratings, financial ratings, and overall ratings. The star ratings are calculated based on a continuous performance score ranging 0 to 100. Accountability and transparency ratings are available since 2011 – when they were introduced; financial ratings are available since at least 2002. A few charities with very serious concerns (like serious accusations or undergoing government investigations) receive Donor Advisory instead of ratings. In addition to discrete star ratings and continuous performance scores, the performance scorecard for each rated charity contains information on all the performance metrics that are used to calculate charity ratings. For instance, the financial performance section shows the data for a charity’s program expenses, administrative expenses, fundraising expenses, fundraising efficiency, working capital, and primary revenue and program expenses growth.

4.4 Methodology

This section outlines how the charity-level data are analyzed to determine whether the performance scores and their components reported by Charity Navigator improve due to the release of charity ratings. Given that Charity Navigator, not a charity by itself, decides which organizations to rate based on explicit eligibility criteria, first issuance of the rating can be treated as a plausibly exogenous shock for the rated organization¹⁵. The Charity Navigator writes about the charities that they evaluate the following: “we are able to evaluate charities with or without their participation.”¹⁶ Even if a charity has been informed about the upcoming release of its rating, it is unlikely that it can undertake actions directed at affecting the results of the evaluation: a rating is calculated based on historical data a charity reports to IRS for a completed fiscal year. The Charity Navigator writes that it can obtain the Form 990 two to three months after it is filed; charities, in turn, have 135 days following the end of a fiscal year to file and often ask for extensions.¹⁷ Also, charities cannot opt out of being rated.¹⁸ As a result, following the issuance of the first rating, according to the theory, a rated organization can learn about its absolute and relative rating status and undertake rating improvement efforts. This research proposal takes advantage of this exogenous shock at the first issuance of the rating and examines whether it is followed by subsequent improvements in performance measures incorporated in the rating system.

¹⁵ Charity Selection Criteria

<http://www.charitynavigator.org/index.cfm?bay=content.view&cpid=32#.VamQKPIViko>

¹⁶ <http://help.charitynavigator.org/kb/questions-about-the-charities-we-rate/can-i-request-to-have-my-charity-removed-from-the-site>

¹⁷ <http://www.charitynavigator.org/index.cfm?bay=content.view&cpid=441#.VamZAvlViko>

¹⁸ <http://help.charitynavigator.org/kb/questions-about-the-charities-we-rate/can-i-request-to-have-my-charity-removed-from-the-site>

To test the proposed model, the analysis is conducted on a panel of charities rated based on their financial reports for fiscal years ended (FYE) 2001 to 2018. The time series includes two periods of data: fiscal years before the calendar year when the first rating was released (*prerating period* coded “0”) and the fiscal years starting the year when the first rating was released and organizational response became possible (rating period, coded “1”). The coding scheme for the main explanatory variable labeled “*CN Rated*” is presented in Table 2. Because the rating agency evaluates charities using their past financial reports, there is a lag between the calendar year the first rating was released, and the fiscal year based on which it was released. In other words, a first-time rating issued to a charity in 2003 can be a grade for the FYE 2001 (two-year lag), which means that there is nothing the charity could do to improve its ratings for the FYE 2001-2002. Also, given that Charity Navigator started releasing its Financial Ratings in 2002 and its Accountability and Transparency ratings in 2011, separate explanatory variables (“*CN Rated*” and “*CN Rated Accountability & Transparency*”) indicating the rating periods for the two measures are used in the analysis. The coding approach is presented in Table 4.2.

Table 4. 2: Coding scheme for time periods in the ratings dataset

First Rating Release Year	Y	Y+1	Y+2	Y+3	Y+i
Fiscal Year (FYE)	Y-n	Y+1-n	Y+2-n	Y+3-n	Y+i-n
First Fiscal Year (FYE) Response Possible			Y(n=i)		
CN Rated Variable Coding	Prerating period (0)		Rating period (1)		
	<i>Example (n = 2)</i>				
First Rating Release Year	2003	2004	2005	2006	2007
Fiscal Year (FYE)	2001	2002	2003	2004	2005
First Fiscal Year (FYE) Response Possible			2003		
	Prerating period (FYE 2001-2002)		Rating period (FYE 2003 - 2005)		

First hypothesis (H1) determines whether performance scores and their components improve after the release of performance ratings for the first time. Similarly to Jin and Leslie (2003), the following estimating equation is used to test the hypothesis:

$$\text{PERF.SCORE}_{it} = \beta_1 \text{CNRated}_{it} + \beta_2 X_{it} + \alpha_i + \gamma_t + e_{it} \quad (1)$$

In this equation:

PERF.SCORE_{it} – performance scores for the charity i at time t

CNRated_{it} – coded “1” for the fiscal years after a charity was first rated (rating period)

Reference group – the years in the prerating period

X_{it} - the vector of control variables

α_i – organization fixed effects

γ_t – year fixed effects

As the theory suggest that charities will respond by improving on the measured performance metrics, PERF.SCORE_{it} is operationalized using each of the following variables included in the CN report cards:

- Overall score
- Financial Score
- Accountability and Transparency Score
- Program Expenses (percent of total expenses)
- Administrative Expenses
- Program Expenses Growth (percent)
- Fundraising Efficiency
- Working Capital Ratio

As in the analysis conducted by Jin and Leslie (2003), there is no unrated control group in this model and the effects of issuing external ratings is estimated relying on time series

variation with year fixed effects capturing year-specific changes that would also be expected in unrated organizations.

Hypotheses H2-H3 test whether the improvement after issuance of ratings varies for charities initially rated differently. For H2, the differences in effects are captured by interacting the variable indicating the rating period with the variable indicating the initial rating a charity received. The hypotheses are tested using the following specification:

$$\text{PERF.SCORE}_{it} = \beta_1 \text{CNRated}_{it} + \beta_2 * (\text{CNRated}_{it} * \text{Init.Rated.Poor}_i) + \beta_3 * (\text{CNRated}_{it} * \text{Init.Rated.NeedsImpr}_i) + \beta_4 X_{it} + \alpha_i + \gamma_i + e_{it} \quad (2)$$

In this equation, Init.Rated.Poor_i , $\text{Init.Rated.NeedsImpr}_i$, and Init.Rated.Good_i are dummy variables created to distinguish between charities initially rated “Poor” (0-1 stars), “needs Improvement” (2 stars), and “Good” (3 stars) by the rater. For instance, the variable Init.Rated.Poor_i is coded “1” for charities that received the grade “Poor” when they were rated for the first time. The group initially rated “Good” is set as the hypothetically least responsive reference category. The agencies that receive the initial rating *Excellent* (4 stars) are excluded from this analysis. For testing H3, the specification from (1) is used on the sub-sample restricted to only the charities that received the excellent rating at the time they became rated.

The differences in the effects across fields with different levels of competition described in H4.1 are estimated by interacting the indicator of a charity being rated with a set of indicators of the competition category it belongs to. There are a few ways to operationalize the extent of competition in the nonprofit sector. One way is to measure the number of nonprofits in a defined group competing for a charitable dollar. The operational measure is the number of public charities in a category normalized by the size of the market in dollars. Charities can be distinguished by categories using the Charity Navigator’s own categorization.

A second way to approach an organization’s motivation to compete for resources is through revenue concentration. According to Trussel & Parsons (2007), “A firm that is dependent on one or a few revenue providers is vulnerable to declines in the economic health or changes in the donation preferences of those providers” (p.269). Research shows this measure predicts nonprofit organization financial vulnerability. Therefore, the fewer revenue sources a charity has, the more motivated it will be to compete for those sources and, therefore, improve its charity ratings. The operational definition of the measure is the sum of the squared shares of each revenue source out of total revenue.

To make the coefficients meaningful in the context of the hypothesis, the reference group in the indicator of the nonprofit category was set to the category with the lowest value on the calculated field competition, and the categories in the factor variable were arranged in the order of increasing competition.

$$\text{PERF.SCORE}_{it} = \beta_1 \text{CNRated}_{it} + \beta_2 * (\text{CNRated}_{it} * \text{Category}_i) + \beta_3 X_{it} + \alpha_i + \gamma_i + e_{it} \quad (3)$$

Finally, H4.2 is estimated by interacting the main explanatory variable with one of four dummies indicating a charity’s share of public contributions in its revenue portfolio. The indicator’s levels correspond to the quartiles in the distribution of the shares of public contributions in the population of rated charities, with the lowest quartile (Q1) representing the reference category:

$$\text{PERF.SCORE}_{it} = \beta_1 \text{CNRated}_{it} + \beta_2 * (\text{CNRated}_{it} * \text{Share.Contrib.Q2}) + \beta_3 * (\text{CNRated}_{it} * \text{Share.Contrib.Q3}) + \beta_4 * (\text{CNRated}_{it} * \text{Share.Contrib.Q4}) + \beta_4 X_{it} + \alpha_i + \gamma_i + e_{it} \quad (4)$$

This analysis controls for a number of factors that can affect financial performance scores, transparency, and accountability. Time-invariant organizational characteristics (such as

corporate culture) and unobserved time-variant environmental factors that affect all charities are controlled by including charity-level fixed effects and year fixed effects accordingly (Chatterji & Toffel, 2010; Lewis, 2014). Additional time-variable factors are also incorporated in the analysis following the insight from (Saxton & Guo, 2011; Trussel & Parsons, 2007). These include Total Revenue and Organizational size. The latter is operationalized as the size of current assets and reflects reputation.

4.5 Findings

Table 4.3 below presents the results of testing hypothesis **H1** for the nine response variables of interest, including the overall performance score, financial score, accountability and transparency score, and five financial metrics that make up the financial score. According to the theory, nonprofit agencies are expected to improve their composite performance scores during the years after they became rated, by improving on at least some of the variables that determine those scores. The analysis of the available data, however, presents results that are contrary to the expectations. First, the coefficient on *CN Rated Financial* variable indicating the fiscal years for which an agency received its financial ratings and could react to them is statistically insignificant and is close to zero even in the sample. The finding suggests that during the years following the issuance of the first charity rating, rated nonprofits, on average, did not improve their financial scores.

Table 4. 3: Agency measured performance after initial Charity Navigator's rating

	1	2	3	4	5	6	7	8	9
	Overall Score	Overall Score	Financial Score	Account. & Transp. Score	Program Expenses (Percent)	Administrative Expenses (Percent)	Fundraising Efficiency	Program Expenses Growth (Percent)	Working Capital Ratio
CN Rated Financial	0.27		-0.21		0.45 *	-0.26	0.02	-4.73 ***	-0.06
CN Rated Accountability	0.25		0.28		0.21	0.15	0.02	0.66	0.05
Total Revenue	0.01 ***	0.01 ***	0.01 ***	-0.00	0.00	-0.00	-0.00	0.02 ***	-0.00 ***
Assets	-0.00***	-0.00***	-0.00***	-0.01*	-0.00***	0.00***	0.00	-0.00	0.00***
Agency Fixed Effects	Included	Included	Included	Included	Included	Included	Included	Included	Included
FYE [2001]	0.59	0.59	0.31		-0.78	0.09	0.02	-1.29	0.33
FYE [2002]	1.10	1.10	1.25		0.92	0.68	0.09	2.90	0.21
FYE [2003]	0.86	1.01	0.91		0.21	-0.26	0.00	-0.99	0.15
FYE [2004]	1.08	1.07	1.23		0.90	0.67	0.09	2.84	0.20
FYE [2005]	-2.05 *	-1.89	-2.06		-0.32	-0.30	-0.01	-3.39	0.31
FYE [2006]	1.05	1.03	1.19		0.87	0.65	0.09	2.76	0.20
FYE [2007]	-2.49 *	-2.30 *	-2.43 *		-0.30	-0.52	-0.01	-3.67	0.21
FYE [2008]	1.04	1.02	1.19		0.87	0.65	0.09	2.74	0.19
FYE [2009]	-2.06 *	-1.79	-1.92		0.07	-1.00	-0.02	-4.39	0.33
FYE [2010]	1.05	1.02	1.20		0.88	0.65	0.09	2.78	0.20
FYE [2011]	-0.79	-0.52	-0.66		0.08	-0.69	-0.03	-3.50	0.39 *
FYE [2012]	1.05	1.02	1.20		0.88	0.65	0.09	2.77	0.20
FYE [2013]	-0.39	-0.11	-0.30		0.67	-1.14	-0.03	-2.31	0.36
FYE [2014]	1.05	1.02	1.20		0.88	0.65	0.09	2.77	0.20
FYE [2015]	-0.09	0.19	0.03		0.90	-1.24	0.01	-1.82	0.43 *
FYE [2016]	1.05	1.02	1.20		0.88	0.65	0.09	2.77	0.20
FYE [2017]	-1.43	-1.14	-1.09	0.36	0.80	-1.11	0.07	-5.96 *	0.79 ***
(Intercept)	1.04	1.00	1.18	3.90	0.86	0.64	0.09	2.73	0.19
	-0.08	0.20	-1.24	3.22	1.32	-1.32 *	0.01	-8.81 **	0.91 ***
	1.04	1.00	1.18	3.90	0.87	0.65	0.09	2.74	0.19
	0.75	0.08	-1.12	2.92	1.28	-1.31 *	-0.03	-10.31 ***	0.95 ***
	1.03	1.05	1.18	3.90	0.86	0.64	0.09	2.73	0.19
	1.51	0.84	-0.39	3.64	1.61	-1.65 *	-0.02	-10.30 ***	0.88 ***
	1.03	1.05	1.18	3.90	0.86	0.64	0.08	2.72	0.19
	1.69	0.94	-0.24	3.55	1.68	-1.83 **	-0.02	-8.45 **	0.92 ***
	1.03	1.05	1.18	3.91	0.86	0.64	0.09	2.73	0.19
	2.34 *	1.50	0.58	3.31	1.87 *	-1.73 **	-0.04	-7.68 **	0.98 ***
	1.03	1.06	1.18	3.91	0.86	0.64	0.09	2.72	0.19
	3.50 ***	2.65 *	1.79	4.50	1.88 *	-1.85 **	-0.03	-7.91 **	0.97 ***
	1.04	1.07	1.19	3.91	0.87	0.65	0.09	2.75	0.20
	3.76 ***	2.86 **	1.82	4.98	1.84 *	-1.78 **	-0.03	-7.84 **	0.94 ***
	1.05	1.08	1.19	3.91	0.88	0.65	0.09	2.76	0.20
	3.71 ***	2.79 *	1.41	5.07	1.67	-1.94 **	-0.02	-9.16 **	1.01 ***
	1.07	1.10	1.22	3.92	0.90	0.67	0.09	2.83	0.20
	78.39***	78.36***	88.71***	64.46***	84.42***	11.92***	0.06	14.86**	1.55***
Observations	197	197	225	423	165	123	0.16	5.20	0.37
R2 / R2 adjusted	10241	10241	10241	6944	10241	10233	10241	10170	10241
	0.522 /	0.523 /	0.502 /	0.734 /	0.739 /	0.707 / 0.679	0.210 / 0.135	0.261 /	0.804 /
	0.477	0.478	0.456	0.696	0.715			0.190	0.786

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Second, even though the coefficient on *CN Rated Accountability* indicating the fiscal years following the year when accountability and transparency ratings became publicly available for a particular charity is, as expected, positive and highly significant, the magnitude of the improvement in the accountability score based on the obtained point estimate (2.43 points with $se = 0.29$ points) has little practical significance. The difference between neighboring star-grades

is 10 points, and there were only 7.26 percent of observations in the population of rated charities (as of 2018) with financial scores between 88 and 90 – the cases where 2.43-point increase in accountability and transparency score would improve the overall star-grade by one star.

Similarly, the positive and significant coefficient on *Program Expenses (percent)* indicates a statistically significant at five percent increase in the fraction of charities' total budget spent on their program activities after they became publicly rated, but the point estimate of 0.45 percent improvement has little practical significance, given the variation in this variable ranging from 10 percent to literally 100 percent and the median of 81.30 percent (see Figure C.1 and Table C.1 in the Appendix C).

In addition to that, a charity's annual program expenses growth rate shrank by the estimated 4.73 percentage points, which is close to the median Program Expenses Grow (see Figure C.2 and Table C.2 in the Appendix C), so the *Overall Score* during the rating period does not become statistically different from the preparing period.

Hypothesis H2 posited that initially rated “Poor” agencies would demonstrate higher levels of improvement than those initially rated good or average. The differences in charity performance changes during the rating period depending on the level of the initial rating are reflected by the coefficients on the interaction terms between *CN Rated Financial / CN Rated Accountability* and the indicator of the initial rating for each agency *Initial Rating* coded as [Poor] (0-1 stars) / [Needs Improvement] (2 stars) / [Good] (3 stars). The agencies that receive the initial rating *Excellent* (4 stars) are excluded from this analysis. As the distribution of initial ratings provided in the Appendix C (Figure C. 3 and Table C. 3) show, the most frequent initial rating is three stars [Good] accounting for 41.53 percent of all initial ratings; poor initial ratings, on the other hand, represent only 9.07 percent in the population of CN rated agencies as of 2018.

The regression results are presented in Table 4.4. The coefficients on *CN Rated Financial* show the performance scores for the reference group (the charities initially rated “Good”) decreased after becoming rated relative to the prerating period by the estimated 1.24 points on the *Overall Score* and 1.54 points on the *Financial Score*. Both changes are statistically significant, although appear to be negligible from the perspective of practical significance. The observed decrease in performance appears to be driven by a slight decrease in *Program Expenses* and a significant slowdown in the annualized *Program Expenses Growth*. The *Accountability and Transparency Score*, however, increased by the estimated 2.20 (0.36) points.

In contrast to the reference group that showed minuscule negative changes in the measured performance indicator, the response for the groups that received lower initial ratings appears to be significantly stronger and in line with the laid-out theory. In the first model, which estimates the changes in the *Overall Score* after an agency becomes publicly rated, both interaction terms estimating the differences in the coefficients for initially low-rated agencies relative to initially highly rated agencies are statistically significant, positive, and also practically significant. Unlike the initially rated “Good” (three stars) reference group that did not show any meaningful improvement after becoming publicly rated, the group that initially received the label “Needs Improvement” (two stars) added the estimated 5.03 points to the difference in the expected response in performance scores of the reference group. As predicted by the theory, the strongest response is observed in the group initially rated “Poor” (0-1 star) with the estimated difference in the coefficients of highly significant 10.26 points. The response of such magnitude is enough to move a charity one step up on the star scale and thus leave its initial grade-category.

Table 4. 4: Agency measured performance after initial Charity Navigator’s rating depending on the value of the initial rating (reference group for Initial Rating = “Good”)

	1 Overall Score	2 Financial Score	3 Account. & Transp. Score	4 Program Expenses (Percent)	5 Administrative Expenses (Percent)	6 Fundraising Efficiency	7 Program Expenses Growth (Percent)	8 Working Capital Ratio
CN Rated	-1.24 ***	-1.54 ***		-0.62 *	0.43	0.02 ***	-6.26 ***	-0.01
Financial	0.35	0.40		0.31	0.24	0.00	0.98	0.06
CN.Rated.Fin *	10.26 ***	9.29 ***		6.67 ***	-3.66 ***	-0.05 ***	10.35 ***	0.08
Initial.Rating.[Poor]	0.70	0.80		0.62	0.47	0.01	1.99	0.11
CN.Rated.Fin *	5.03 ***	4.86 ***		2.01 ***	-1.77 ***	-0.02 *	8.39 ***	-0.04
Initial.Rating.[NeedsImpr]	0.50	0.57		0.45	0.34	0.01	1.41	0.08
CN Rated			2.20 ***					
Accountability			0.36					
CN.Rated.Acc *			2.31 ***					
Initial.Rating.[Poor]			0.60					
Rated.Acc *			1.74 ***					
Initial.Rating.[NeedsImpr]			0.42					
Total Revenue	0.01 **	0.01 *	-0.00	-0.00	0.00	-0.00	0.02 ***	-0.00 *
	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
Assets	0.03 ***	0.05 ***	-0.03 **	0.02 ***	-0.01 ***	-0.00 *	0.08 ***	0.01 ***
	0.01	0.01	0.01	0.01	0.00	0.00	0.02	0.00
Agency Fixed Effects	Included	Included	Included	Included	Included	Included	Included	Included
FYE [2001]	1.66	1.22		-0.84	-0.43	0.03	6.06	-0.07
	1.39	1.60		1.24	0.95	0.02	3.92	0.22
FYE [2002]	2.43	2.28		0.44	-0.69	0.02	7.57 *	-0.25
	1.36	1.57		1.22	0.93	0.02	3.83	0.22
FYE [2003]	-0.35	-0.75		-0.04	-0.86	0.02	6.21	-0.30
	1.32	1.52		1.18	0.90	0.02	3.72	0.21
FYE [2004]	-0.25	-0.69		-0.30	-0.80	0.02	5.14	-0.24
	1.31	1.51		1.17	0.89	0.02	3.70	0.21
FYE [2005]	-0.43	-0.78		0.18	-1.53	0.02	3.58	-0.17
	1.32	1.53		1.19	0.90	0.02	3.74	0.21
FYE [2006]	1.65	1.26		0.61	-1.33	0.01	4.36	-0.07
	1.32	1.53		1.19	0.90	0.02	3.73	0.21
FYE [2007]	1.48	1.04		1.09	-1.76	0.01	5.65	-0.23
	1.33	1.53		1.19	0.90	0.02	3.74	0.21
FYE [2008]	1.17	0.76		1.28	-1.69	0.02	5.63	-0.08
	1.32	1.53		1.19	0.90	0.02	3.74	0.21
FYE [2009]	1.08	0.62	0.63	1.67	-1.89 *	0.01	1.10	0.27
	1.30	1.51	3.93	1.17	0.89	0.02	3.68	0.21
FYE [2010]	2.19	0.11	3.09	2.32 *	-2.07 *	0.01	-1.78	0.34
	1.31	1.51	3.92	1.17	0.89	0.02	3.69	0.21
FYE [2011]	3.50 **	1.04	2.14	2.42 *	-2.23 *	0.00	-2.78	0.43 *
	1.30	1.51	3.93	1.17	0.89	0.02	3.68	0.21
FYE [2012]	4.41 ***	1.90	2.88	2.77 *	-2.48 **	0.01	-2.09	0.33
	1.30	1.50	3.93	1.16	0.89	0.02	3.66	0.21
FYE [2013]	4.65 ***	2.19	2.86	2.86 *	-2.68 **	0.01	-0.25	0.39
	1.30	1.51	3.93	1.17	0.89	0.02	3.68	0.21
FYE [2014]	5.34 ***	2.84	2.71	3.14 **	-2.61 **	0.00	0.74	0.43 *
	1.30	1.50	3.93	1.17	0.89	0.02	3.67	0.21
FYE [2015]	6.54 ***	4.10 **	4.03	3.15 **	-2.76 **	-0.00	0.40	0.45 *
	1.31	1.52	3.93	1.18	0.90	0.02	3.71	0.21
FYE [2016]	6.79 ***	4.06 **	4.50	3.17 **	-2.70 **	-0.00	0.14	0.40
	1.32	1.52	3.94	1.18	0.90	0.02	3.72	0.21
FYE [2017]	7.06 ***	3.92 *	4.91	2.92 *	-2.87 **	0.00	-0.95	0.46 *
	1.35	1.55	3.94	1.21	0.92	0.02	3.80	0.22
(Intercept)	77.06 ***	87.84 ***	65.24 ***	84.31 ***	12.13 ***	0.04	8.16	2.03 ***
	2.09	2.41	4.26	1.87	1.42	0.03	5.89	0.34
Observations	7190	7190	5002	7190	7182	7190	7136	7190
R ² / R ² adjusted	0.560 /	0.526 /	0.760 /	0.730 /	0.701 / 0.671	0.537 / 0.491	0.261 /	0.807 /
	0.515	0.478	0.724	0.703			0.186	0.788

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

The coefficients on interaction terms for *Accountability and Transparency Score* show similar patterns but with substantially smaller magnitudes. Whereas the estimated improvement in the group initially rated “Good” is estimated be 2.20 points, the group rated “Poor” added estimated 2.31 points to the expected difference. This analysis shows that the described improvements in the *Overall Score* are driven largely by improvements in the *Financial Score*, which, in turn, results from a significant increase in the reported *Program Expenses* and *Program Expenses Growth*, as well as a cutback in *Administrative Expenses*. Overall, the findings confirm the hypothesis with respect to the *Overall Performance* score including its both components.

Testing hypothesis H3, which argues that charities that initially receive the highest rating will subsequently reduce their measured performance, is based on a subsample restricted to the charities that initially received a four-star overall performance grade with the “Excellent” label. As hypothesized, the analysis of the data shows a statistically significant decrease in the expected *Overall Score* by the estimated 3.87 points after a charity becomes rated (Table 4.5). This change reflects a statistically significant drop of comparable magnitude in the expected *Financial Score* and no change in the *Accountability and Transparency* score. The change in the *Financial Score* appears to be driven by a significant slowdown in the annualized *Program Expenses Growth* by the estimated expected 11.72 percentage points.

Table 4. 5: Agency measured performance after initial Charity Navigator’s rating for initially top-rated charities

	1	2	3	4	5	6	7	8
	Overall Score	Financial Score	Account. & Transp. Score	Program Expenses (Percent)	Administrative Expenses (Percent)	Fundraising Efficiency	Program Expenses Growth (Percent)	Working Capital Ratio
CN Rated	-3.87***	-4.16***		-0.76*	0.72**	0.06	-11.72***	-0.15
Financial	0.45	0.52		0.33	0.24	0.07	1.11	0.12
CN Rated			-0.46					
Accountability			0.66					
Total Revenue (Mil. Dollars)	0.02***	0.03***	-0.01	0.00	0.00	-0.00	0.02*	-0.00***
Assets (Mil. Dollars)	-0.00***	-0.01***	-0.00	-0.00*	0.00**	0.00	-0.00	0.00***
Agency Fixed Effects	Included	Included	Included	Included	Included	Included	Included	Included
FYE [2001]	0.04	-0.05		-0.40	0.64	-0.00	-10.52**	0.85*
	1.61	1.83		1.17	0.84	0.26	3.92	0.42
FYE [2002]	-0.06	0.14		0.01	0.11	-0.03	-11.02**	0.69
	1.57	1.79		1.15	0.82	0.25	3.84	0.41
FYE [2003]	-2.28	-2.08		-0.18	0.10	-0.09	-15.02***	1.16**
	1.54	1.75		1.12	0.80	0.25	3.76	0.40
FYE [2004]	-3.70*	-3.34		0.38	-0.66	-0.07	-13.85***	0.76
	1.54	1.75		1.12	0.80	0.25	3.75	0.40
FYE [2005]	-2.84	-2.54		0.00	-0.36	-0.11	-14.07***	0.95*
	1.56	1.78		1.14	0.81	0.25	3.81	0.41
FYE [2006]	-3.28*	-3.01		-0.94	0.27	-0.11	-13.09***	0.91*
	1.56	1.78		1.14	0.81	0.25	3.81	0.41
FYE [2007]	-2.20	-1.99		-0.41	-0.04	-0.10	-12.78***	1.11**
	1.56	1.78		1.14	0.81	0.25	3.81	0.41
FYE [2008]	-0.63	-0.38		-0.02	-0.55	-0.02	-11.13**	1.06**
	1.56	1.78		1.14	0.81	0.25	3.81	0.41
FYE [2009]	-4.12**	-3.07		-0.96	0.16	0.19	-13.93***	1.44***
	1.53	1.75		1.12	0.80	0.25	3.75	0.40
FYE [2010]	-2.38	-2.62	3.72***	-0.81	-0.08	-0.00	-16.78***	1.64***
	1.54	1.76	0.66	1.13	0.81	0.25	3.77	0.41
FYE [2011]	-2.33	-4.08*	6.25***	-1.09	0.27	-0.09	-19.10***	1.57***
	1.54	1.76	0.88	1.12	0.80	0.25	3.77	0.41
FYE [2012]	-2.05	-3.74*	6.98***	-0.91	-0.20	-0.09	-20.67***	1.57***
	1.54	1.76	0.89	1.12	0.80	0.25	3.77	0.41
FYE [2013]	-2.34	-4.43*	6.70***	-1.02	-0.21	-0.09	-19.62***	1.54***
	1.54	1.76	0.91	1.13	0.81	0.25	3.78	0.41
FYE [2014]	-1.77	-3.22	6.31***	-1.04	-0.06	-0.11	-19.67***	1.62***
	1.53	1.74	0.91	1.12	0.80	0.25	3.74	0.40
FYE [2015]	-0.84	-2.28	7.18***	-1.09	-0.05	-0.10	-19.84***	1.53***
	1.55	1.77	0.94	1.13	0.81	0.25	3.79	0.41
FYE [2016]	-0.39	-2.00	7.75***	-1.21	0.00	-0.10	-18.81***	1.53***
	1.56	1.78	0.96	1.14	0.81	0.25	3.82	0.41
FYE [2017]	-1.20	-3.09	7.09***	-1.15	-0.17	-0.07	-20.86***	1.63***
	1.61	1.84	1.03	1.18	0.84	0.26	3.94	0.42
(Intercept)	93.57***	93.35***	88.51***	75.78***	14.90***	0.10	41.87***	0.12
	1.93	2.20	1.51	1.41	1.01	0.31	4.72	0.51
Observations	3051	3051	1942	3051	3051	3051	3034	3051
R ² / R ² adjusted	0.405 / 0.352	0.438 / 0.388	0.564 / 0.503	0.756 / 0.735	0.730 / 0.706	0.204 / 0.133	0.321 / 0.261	0.804 / 0.787

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

The results of testing the last two hypotheses that address the moderating effect of competition for donated revenue are provided in Tables 4.7 and 4.8. Hypothesis H4.1 posits that

nonprofit agencies in fields with more competition will demonstrate a greater level of improvement in response to ratings than organizations in markets with less competition. The variable measuring the extend of market competition among rated charities was computed as the average over the fiscal years 2000 – 2018 number of public charities in a category normalized by the size of the market in dollars in that category. Table 4.6 presents the values on the measure for the different charity categories arranged in the descending order. To test the hypothesis, the binary variables indicating each charitable category were included in the regression as part of the interaction term with the indicator of the period during which a charity was rated. The results of the multiple regression analysis are presented in Table 4.7. The “International” category showing the lowest value on the competition measure is the reference group and the regression coefficients are listed in the order of increasing competition.

Table 4. 6: Computed average competition by category of charitable activity (sorted from the highest competition to the lowest)

Category	Average Competition (2000-2018)
Human and Civil Rights	134.30
Environment	123.66
Religion	91.00
Health	87.68
Animals	81.85
Education	74.73
Arts, Culture, Humanities	70.50
Research and Public Policy	58.08
Human Services	55.77
Community Development	48.51
International	31.63

Table 4. 7: Moderating effect of the competition across fields of activity

	1	2	3	4	5	6	7	8
	Overall Score	Financial Score	Account. & Transp. Score	Program Expenses (Percent)	Administrative Expenses (Percent)	Fundraising Efficiency	Program Expenses Growth (Percent)	Working Capital Ratio
CN Rated Financial	-0.23	-0.88		0.09	0.89 *	0.01	-11.58 ***	-0.05
	0.67	0.77		0.56	0.42	0.06	1.78	0.13
CN Rated Financial *	0.98	0.60		0.55	-1.31 *	0.01	6.58 **	0.02
Community	0.93	1.06		0.78	0.58	0.08	2.45	0.17
Development								
FinPostRated1 *	0.96	1.35		-0.04	-0.81	0.00	9.34 ***	0.07
Human Services	0.75	0.86		0.63	0.47	0.06	1.99	0.14
FinPostRated1 *	-5.38 ***	-5.23 **		-0.56	-0.33	-0.00	1.78	-0.43
Research.&.Public Policy	1.46	1.66		1.22	0.91	0.12	3.83	0.27
FinPostRated1 *	0.95	1.43		1.22	-2.39 ***	0.01	8.34 ***	-0.14
Arts, Culture, Humanities	0.83	0.95		0.69	0.52	0.07	2.20	0.16
FinPostRated1 *	1.07	0.93		-0.21	-0.73	-0.00	7.69 **	0.20
Education	1.01	1.16		0.85	0.63	0.08	2.68	0.19
FinPostRated1 *	-0.37	0.61		1.12	-0.90	0.00	-13.68 ***	0.10
Animals	1.10	1.25		0.92	0.68	0.09	2.89	0.21
FinPostRated1 *	-1.16	-1.55		-0.79	-0.50	0.11	7.25 **	-0.00
Health	0.86	0.98		0.72	0.54	0.07	2.28	0.16
FinPostRated1 *	1.78	3.69 **		3.45 ***	-3.43 ***	-0.02	10.77 ***	-0.32
Religion	1.08	1.23		0.90	0.67	0.09	2.83	0.20
FinPostRated1 *	1.55	0.76		1.20	-1.13	-0.02	7.84 **	0.01
Environment	1.05	1.19		0.87	0.65	0.09	2.75	0.20
FinPostRated1 *	-0.15	-0.47		0.12	-2.10 **	0.00	5.97	-0.18
Human and Civil Rights	1.25	1.42		1.04	0.78	0.10	3.31	0.23
AccPostRated1			5.62 ***					
			0.59					
AccPostRated1 *			-3.58 ***					
Community			0.74					
Development								
AccPostRated1 *			-3.13 ***					
Human Services			0.59					
AccPostRated1 *			-4.83 ***					
Research and Public Policy			1.18					
AccPostRated1 *			-4.51 ***					
Arts, Culture, Humanities			0.66					
AccPostRated1 *			-2.72 **					
Education			0.84					
AccPostRated1 *			-2.82 **					
Animals			0.89					
AccPostRated1 *			-2.59 ***					
Health			0.68					
AccPostRated1 *			-2.70 **					
Religion			0.86					
AccPostRated1 *			-2.58 **					
Environment			0.84					
AccPostRated1 *			-3.43 ***					
Human & Civil Rights			1.00					
Total Revenue	0.01 ***	0.01 ***	-0.00	0.00	-0.00	-0.00	0.02 ***	-0.00 ***
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Assets	-0.00 ***	-0.00 ***	-0.01 *	-0.00 ***	0.00 ***	0.00	-0.00	0.00 ***
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Agency Fixed Effects	Included	Included	Included	Included	Included	Included	Included	Included
Year Fixed Effects	Included	Included	Included	Included	Included	Included	Included	Included
(Intercept)	78.82 ***	88.48 ***	64.56 ***	83.80 ***	11.68 ***	0.07	30.91 ***	1.42 ***
	2.10	2.39	4.24	1.75	1.31	0.17	5.52	0.39
Observations	10241	10241	6944	10241	10233	10241	10170	10241
R ² / R ² adjusted	0.524 /	0.505 /	0.736 /	0.740 /	0.709 / 0.681	0.210 /	0.269 /	0.805 /
	0.479	0.457	0.698	0.715		0.135	0.198	0.786

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Although the expectation was to see the coefficients on the interacting terms to steadily increase, this does not appear to be what the data show. Except one coefficient, the differences among the categories on the measure of the *Overall Score* appear to be statistically insignificant and close to zero even in the sample. The pattern is similar on the measure of the *Financial Score*. The regression output shows some small but statistically significant differences on the interaction terms for the *Accountability and Transparency* score, but these differences do not reflect the hypothesized expectations. For instance, the last two interaction terms with the categories having the highest values on the measure of competition (Environment and Human and Civil Rights) are expected to have the largest significant coefficients as opposed to the coefficients on the terms at the top of the list. Overall, the analysis of the available data fails to confirm Hypothesis 4.1.

Finally, the results of the regression analysis testing the competition hypothesis H4.2 are presented in Table 4.8. The proposed theory argues that charities with a greater share of public contributions in their revenue portfolios will demonstrate a greater level of improvement in response to ratings than organizations that rely on contributions to a lesser extent. The descriptive analysis of the share of public contributions in the in nonprofit income portfolios is presented in the Appendix C (Figure C4 and Table C4). It shows that the population of charities at the Charity Navigator includes the whole spectrum of agencies from those having practically no public contributions in a given fiscal year to those that rely on them entirely. The distribution is, however, heavily skewed to the left with the median share of public donations in the revenue portfolio being equal to 86 percent.

To conduct the hypothesis test, the numeric values of the share of contributions in each charity's income portfolio was recoded into one of four categories based on the distribution

quartiles and interacted with the indicator of the rating period. The group of organizations that belongs to the lowest quartile on this measure was set as the reference group. Table 6 shows how the response to being rated changes for the charities in the higher quartiles of the distribution relative to the reference group. Contrary the theoretical arguments, all the interaction terms in the model for the *Overall Score* and in the model for the *Financial Score* are near zero and statistically insignificant. In other words, charities across different levels of reliance on charitable contributions in their revenue portfolios appear to be equally unresponsive to being rated in terms of expected improvements in their performance scores.

In the model for *Accountability and Transparency* score, the group of agencies in the top quartile on the studied measure (96-100 percent reliance on charitable contributions) demonstrates the estimated 2.46 point higher improvement in the score in addition to the 1.13 point improvement estimated for the reference group. Still, despite being statistically significant, the responses observed in the accountability and transparency score have quite little substantive value and have minimal impact on the *Overall Score*. Therefore, it can be concluded that the empirical tests find no support for the competition hypothesis 4.2.

Table 4. 8: Moderating effect of the share of public contributions on agency response to ratings

	1	2	3	4	5	6	7	8
	Overall Score	Financial Score	Account. & Transp. Score	Program Expenses (Percent)	Administrative Expenses (Percent)	Fundraising Efficiency	Program Expenses Growth (Percent)	Working Capital Ratio
CN Rated Financial	0.47	0.65		0.73	-0.79 **	0.02	0.87	0.03
	0.47	0.54		0.39	0.29	0.04	1.23	0.08
CN Rated Financial *	-0.60	-1.31		-0.23	0.46	0.01	-8.49 ***	-0.08
Share.Contrib.Q2	0.59	0.67		0.49	0.36	0.05	1.53	0.10
CN Rated Financial *	0.21	-0.33		0.70	-0.28	-0.02	-2.75	-0.19
Share Contrib. Q3	0.61	0.70		0.51	0.38	0.05	1.59	0.11
CN Rated Financial *	-0.24	-1.14		-1.31 **	1.72 ***	-0.02	-7.56 ***	-0.08
Share.Contrib.Q4	0.59	0.67		0.49	0.37	0.05	1.55	0.11
CN Rated A&T			1.13*					
			0.45					
CN Rated A&T *			0.74					
Share Contrib. Q2			0.49					
CN Rated A&T *			0.53					
Share Contrib. Q3			0.50					
CN Rated A&T *			2.46 ***					
Share Contrib. Q4			0.48					
Total Revenue	0.01 ***	0.01 ***	-0.00	0.00	-0.00	-0.00	0.02 ***	-0.00 ***
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Assets	-0.00 ***	-0.00 ***	-0.01 *	-0.00 ***	0.00 ***	0.00	-0.00	0.00 ***
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Agency Fixed Effects	Included	Included	Included	Included	Included	Included	Included	Included
FYE [2001]	0.56	0.23		-1.02	0.31	0.01	-1.38	0.33
	1.11	1.26		0.93	0.69	0.09	2.89	0.20
FYE [2002]	0.88	0.81		0.09	-0.00	-0.00	-1.79	0.30
	1.09	1.24		0.91	0.67	0.09	2.83	0.19
FYE [2003]	-1.81	-1.98		-0.51	-0.14	0.01	-3.84	0.31
	1.06	1.20		0.88	0.65	0.09	2.75	0.19
FYE [2004]	-2.60 *	-2.67 *		-0.44	-0.41	-0.01	-4.84	0.28
	1.05	1.19		0.87	0.65	0.09	2.73	0.19
FYE [2005]	-2.14 *	-2.16		-0.07	-0.88	-0.02	-5.48 *	0.38 *
	1.06	1.21		0.89	0.66	0.09	2.76	0.19
FYE [2006]	-0.86	-0.90		-0.01	-0.57	-0.03	-4.46	0.40 *
	1.06	1.21		0.88	0.66	0.09	2.76	0.19
FYE [2007]	-0.43	-0.51		0.51	-1.02	-0.02	-3.64	0.41 *
	1.06	1.21		0.89	0.66	0.09	2.76	0.19
FYE [2008]	0.06	-0.00		0.79	-1.18	0.01	-3.38	0.53 **
	1.06	1.21		0.89	0.66	0.09	2.76	0.19
FYE [2009]	-1.23	-1.01	0.16	0.76	-1.03	0.08	-6.90 *	0.84 ***
	1.05	1.19	3.90	0.87	0.65	0.09	2.73	0.19
FYE [2010]	-0.14	-1.48	3.11	1.15	-1.21	0.01	-10.01 ***	0.94 ***
	1.05	1.20	3.89	0.88	0.65	0.09	2.73	0.19
FYE [2011]	0.72	-1.34	3.23	1.13	-1.20	-0.02	-11.26 ***	0.98 ***
	1.05	1.19	3.90	0.87	0.65	0.09	2.72	0.19
FYE [2012]	1.43	-0.67	3.81	1.46	-1.53 *	-0.02	-11.31 ***	0.89 ***
	1.04	1.19	3.90	0.87	0.64	0.09	2.71	0.19
FYE [2013]	1.58	-0.45	3.61	1.55	-1.71 **	-0.02	-9.58 ***	0.92 ***
	1.05	1.19	3.90	0.87	0.65	0.09	2.72	0.19
FYE [2014]	2.40 *	0.54	3.48	1.84 *	-1.68 **	-0.04	-8.36 **	0.99 ***
	1.04	1.19	3.90	0.87	0.64	0.09	2.71	0.19
FYE [2015]	3.51 ***	1.62	4.74	1.80 *	-1.81 **	-0.03	-8.65 **	0.98 ***
	1.05	1.20	3.90	0.88	0.65	0.09	2.74	0.19
FYE [2016]	3.69 ***	1.58	5.17	1.65	-1.67 *	-0.03	-8.82 **	0.96 ***
	1.06	1.21	3.90	0.88	0.65	0.09	2.75	0.19
FYE [2017]	3.78 ***	1.34	5.33	1.51	-1.83 **	-0.02	-9.96 ***	1.05 ***
	1.09	1.24	3.91	0.90	0.67	0.09	2.82	0.19
(Intercept)	78.22 ***	88.09 ***	65.46 ***	84.29 ***	12.31 ***	0.06	10.68 *	1.44 ***
	2.00	2.28	4.22	1.67	1.24	0.17	5.20	0.36
Observations	9603	9603	6457	9603	9595	9603	9536	9603
R ² / R ² adjusted	0.520 /	0.499 /	0.743 /	0.736 /	0.709 / 0.680	0.221 / 0.143	0.273 /	0.819 /
	0.472	0.449	0.704	0.709			0.200	0.801

Standard errors in parentheses. * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

4.6 Conclusion

One of the key goals of nonprofit report cards is to improve the efficiency of philanthropic activities. Report cards could increase allocative efficiency both by reallocating donations to more efficient charities (and weeding out inefficient charities) and by getting charities to improve their own performance. By drawing on economic theory of organizational behavior, performance management theory, and institutional theory, this paper proposes a framework predicting that a public charity will respond to an exogenous shock - the release of its charity rating by improving its measured performance, especially if it (1) initially gets a poor rating, (2) is in a highly competitive subfield, (3) relies more heavily on donations.

The empirical tests only partially confirmed the proposed hypotheses. The theory posits that after becoming rated, charities would improve on the measures that affect their public performance scores so that their public performance scores/ratings improve. At the same time, the charities that were top rated would lower their externally measured reported performance. Yet, the analysis of the data, besides confirming the latter argument by finding modest but significant decline in the performance scores of the top rated charities, finds that only the charities that received the lowest two rating grades of 0 and 1-star (labelled by CN “Exceptionally Poor” and “Poor”) meaningfully improved their expected Overall Performance score after becoming publicly rated. The group that initially received a two-star rating labelled “Needs Improvement” showed a statistically significant but modest expected gains in its performance scores. The group that was initially given three stars and labelled “Good”, despite being expected to further improve, demonstrated an immaterial decline in its expected financial and overall performance scores. All the improvements in the expected *Accountability and Transparency Score*, despite their statistical significance, also appear to be only peripheral.

Thereby, only the initial rating has proven to be a significant moderator of a charity's subsequent response. The analysis finds no evidence that competition (operationalized by a measure of industry crowdedness) and the extent of reliance on public contributions, could influence the effects of charity ratings on the behavior of a rated nonprofit organization.

Overall, public charities only respond in a limited way to being publicly rated, meaning limited effectiveness of the existing tool to elicit performance improvements in nonprofits. At the same time, the statistically and practically significant findings for the charities that initially receive the lowest ratings show that nonprofit performance monitoring has some potential. Hence, this research points to some important factors that could potentially explain and influence the observed behaviors of rated charities including informational content of charity ratings, performance standards and thresholds applied, or even reporting and publicizing approaches. An essential continuation of this research would be a further attempt to understand how performance ratings are perceived and reactions formed from within third-sector organizations. These insights might have significant implications for the methodologies used by the performance monitoring community and also for the nonprofit management practice.

4.7 Limitations and Future Research Directions

This study has an exploratory character and several limitations to the significance and generalizability of its findings, which also point to opportunities for further research. The most critical limitations are the following:

First, out of 286,420 public charities that reported to IRS in 2012¹⁹, a relatively small number - only about 9,089 organizations were rated by Charity Navigator as of 2018 and only 8640 received meaningful performance grades. These rated organizations are treated as the population of interest here, and the random sample used in this study was drawn from it. In other words, the results of the inference tests and findings presented in this study are valid for the population of rated charities and their validity is limited outside of the described scope. Due to the limited external validity of this analysis, it is important to admit that charities that are not rated at this time and statistically different from the population of currently rated charities may exhibit different patterns of behavior from those that this analysis uncovered. Also, nonprofit organizations' behaviors in response to charity ratings may change over time as more charities receive ratings and rating methodologies evolve.

Second, the estimates of charity response to being rated relies on time series variation in the performance scores assigned by the rater and their driver-variables due to introduction of ratings. This study does not take advantage of an unrated comparison group, to estimate a stronger, from the perspective of internal validity, difference-in-difference model. Such analysis could potentially be conducted by splitting the sample of rated charities into cohorts based on the year they were first rated, constructing a statistically similar comparison group for each cohort, and, finally, estimating and averaging each cohort's responses to public ratings. Given the findings of this study, however, this step seems to be excessive for all but initially poorly rated nonprofit agencies.

¹⁹ Source: <http://www.urban.org/sites/default/files/alfresco/publication-pdfs/413277-The-Nonprofit-Sector-in-Brief---.PDF>

Third, the analysis of charity response on the measure of accountability and transparency is limited. This study estimates the change in the Accountability and Transparency score, after the charity becomes rated, but does not analyze changes in the variables that determine the score. Unlike is the case with the Financial Score, there is no data available on the accountability and transparency components of the metric to allow such analysis.

Finally, the measure of sub-field competition used in this study is weak, which could explain the findings. Developing a stronger measure leaves room for further research.

APPENDIX A

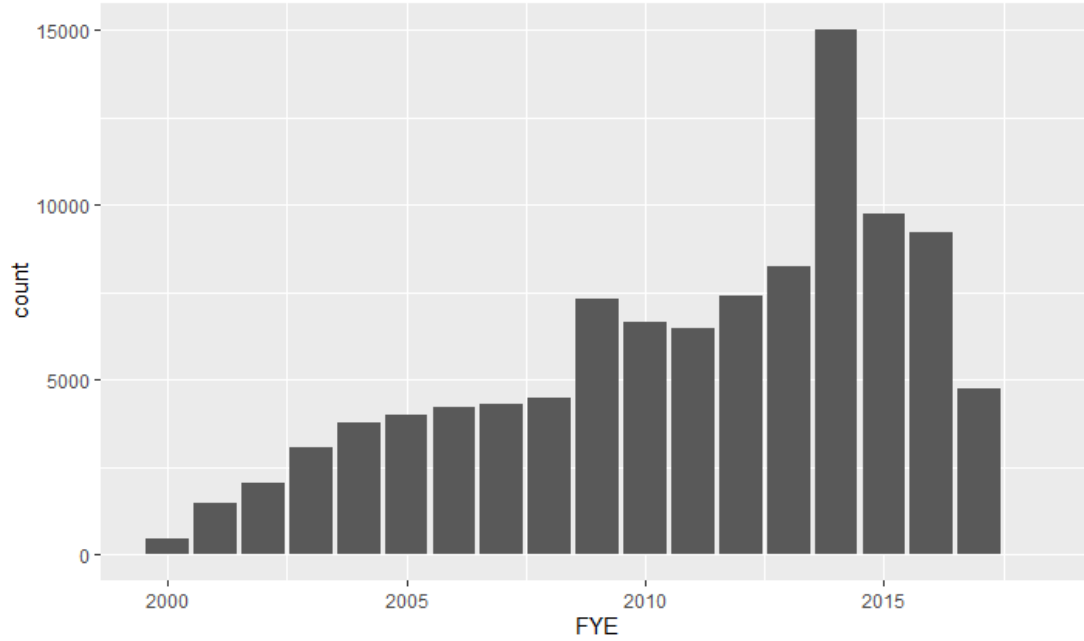


Figure A. 1: The distribution of the 102,534 charity ratings for the 8640 charities rated by the Charity Navigator

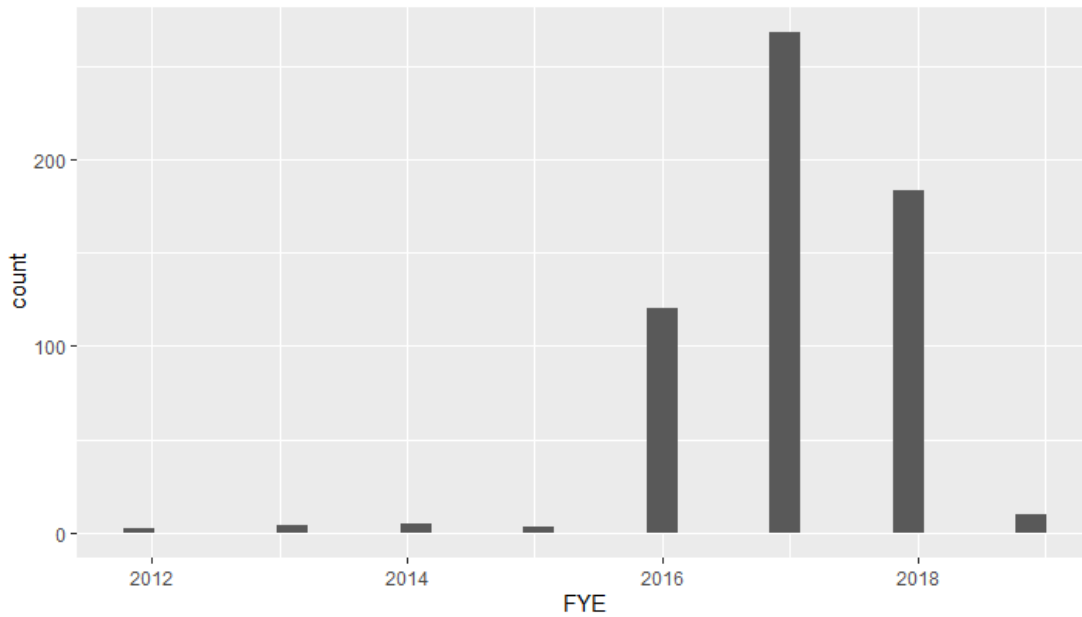


Figure A. 2: The distribution of the 595 charities rated by the Charity Watch

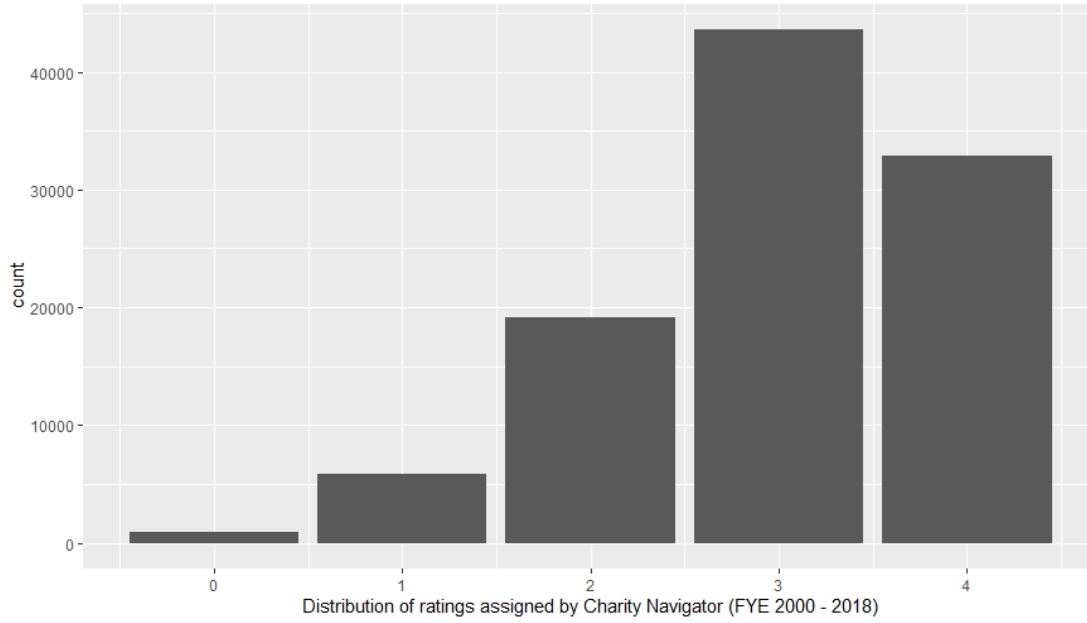


Figure A. 3: The distribution of performance grades assigned by Charity Navigator

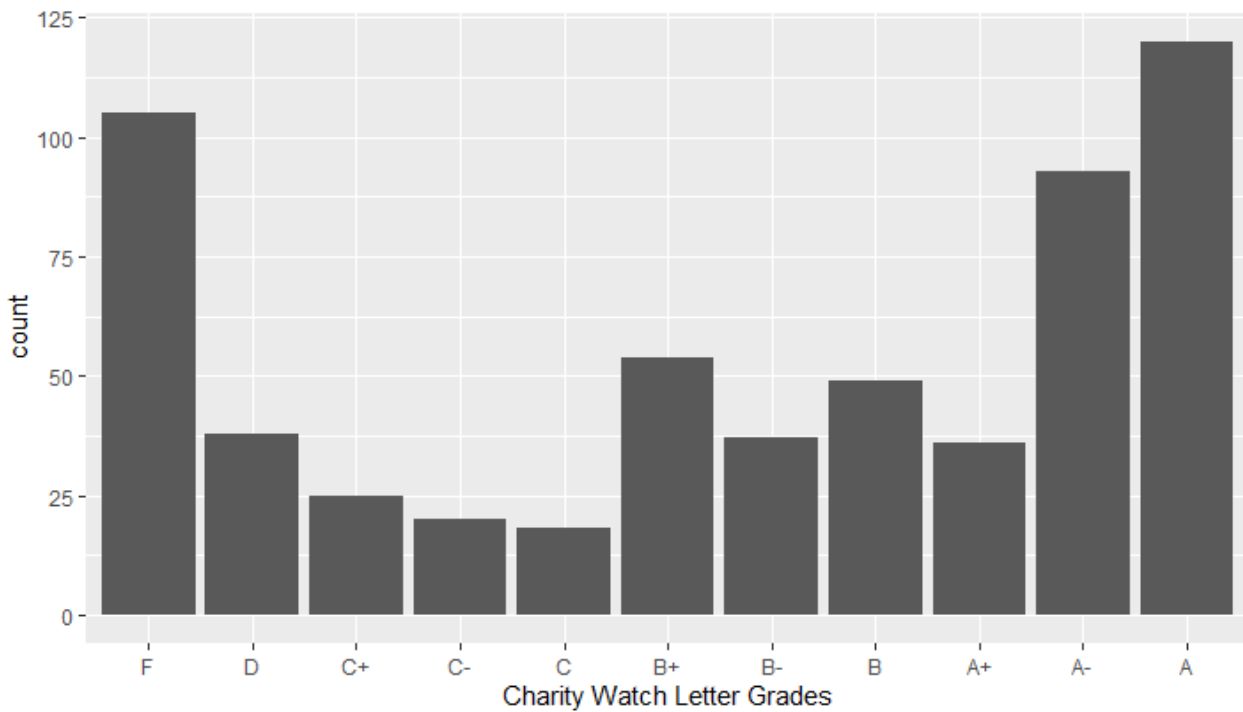


Figure A. 4: The distribution of performance grades assigned by Charity Watch

Table A. 1: Grade Conversion Scheme

Charity Navigator Star-Grades	Charity Navigator Converted Numeric Grades	Charity Watch Letter-based Scale	Charity Watch Converted Numeric Scale (native)	Charity Watch Converted Numeric Scale (adapted to CN)
0 stars	0	F	0	0
1 stars	1	D	1	0
2 stars	2	C-	2	0
3 stars	3	C-	3	1
4 stars	4	C+	4	1
		B-	5	2
		B	6	2
		B+	7	3
		A-	8	3
		A	9	4
		A+	10	4

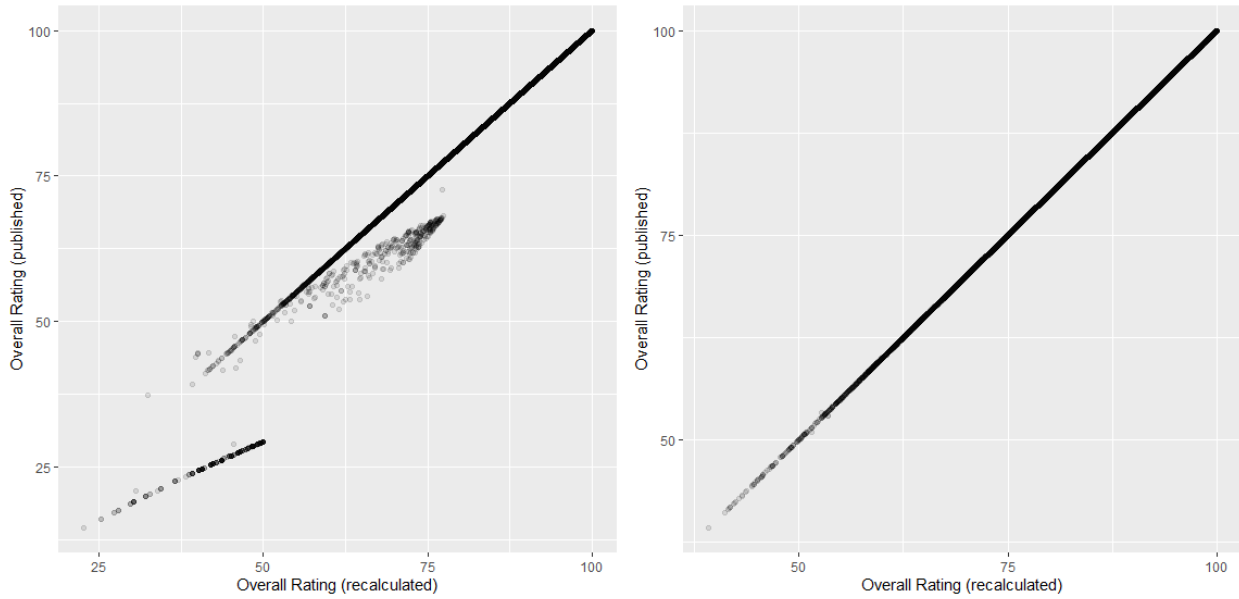


Figure A. 5: Overall published rating scores against overall recalculated scores before and after cleaning the dataset.

APPENDIX B

Table B. 1: Charities selected for the experiment and performance report cards presented



Low Rating & High Program Expenses	High Rating & Low Program Expenses
<p>CHILDREN’S RELIEF MISSION Category: International Cause: Humanitarian Relief Supplies Mission: The organization endeavors to provide clothing, medical supplies, and educational materials to recipients in Third World countries. It also provides cash grants to charitable foreign locations. Self-described accomplishments: Grants/shipments of goods to villagers of various Third World countries. Total Revenue (FYE 2014): \$3,137,634 Program Expenses: (FYE 2014): 99.1 % Overhead expenses: 0.8% The Charity’s Overall Performance Rating:  (Aug 2015)</p>	<p>STAND FOR CHILDREN LEADERSHIP CENTER Category: Education Cause: Education Policy & Reform Mission: To ensure that all children, regardless of their background, graduate from high school prepared for, and with access to, a college education. Self-described accomplishments: Our national programs educate and empower parents, teachers, and community members to demand excellent public schools. We educate the public about, and advocate for, effective state and district level education policies. We ensure that new policies and funding reach classrooms and help students. Our staff provides parents and others concerned about children’s issues with tools to achieve long-lasting improvements for children with one unified voice. Total Revenue (FYE 2014): \$18,537,796 Program Expenses: (FYE 2014): 64.7 % Overhead expenses: 35.2% The Charity’s Overall Performance Rating:  (Sep 2015)</p>

Table B.1 (continued)



<p>GLOBAL SOLUTIONS FOR INFECTIOUS DISEASES Category: Health Cause: Medical Research Mission: We are a non-profit global health organization engaged in the development of diagnostic and preventive tools for infectious diseases, including HIV. We also provide assistance to, and collaborate with, global public health organizations, private foundations, other non-governmental organizations and for-profit entities focused on public health issues and infectious diseases. Self-described accomplishments: Conducting research and developing vaccines and diagnostics for life-threatening infectious diseases, including HIV. Providing assistance to and collaborating with other global public health organizations, private foundations, and for-profit entities focused on public health issues, for the purpose of facilitating access to affordable health solutions which benefit the people in developing countries, who are the most in need. Total Revenue (FYE 2013): \$1,980,374 Program Expenses (FYE 2013): 90.2% Overhead expenses: 9.8% The Charity’s Overall Performance Rating:  (Aug 2015)</p>	<p>FREE TO BREATHE Category: Health Cause: Diseases, Disorders, and Disciplines Mission: To ensure surviving lung cancer is the expectation, not the exception. To turn this vision into reality, we focus on: funding research with the greatest potential to save lives; increasing the number of lung cancer patients participating in clinical trials; building and empowering the lung cancer community. Self-described accomplishments: Providing grants to 12 research institutions across the US; administering the lung cancer mutation consortium; providing funds to member institutions to offset the cost of tumor testing; organizing free to breathe community events in more than forty locations; hosting the lung cancer advocacy summit. Total Revenue (FY 2014): \$3,704,535 Program Expenses (FY 2014): 67.7% Overhead expenses: 32.3% The Charity’s Overall Performance Rating:  (Jun 2016)</p>
---	---

Table B. 2: Demographic characteristics of the study sample by treatment groups

Demographics	Qualtrics sample	T1	T2	T3	T4
n (count)	873.0	208	225	214	226
<i>Gender (%)</i>					
Female	72.6	74.0	75.6	70.1	70.8
Male	27.4	26.0	24.4	29.9	29.2
Age (mean)	43.0	42.5	43.7	42.6	43.3
<i>Age (%)</i>					
Under 25	10.1	12.5	8.9	8.9	10.2
25-34	25.4	24.0	24.0	24.8	28.8
35-44	20.4	20.7	22.2	22.9	15.9
45-54	16.6	16.3	15.6	18.7	15.9
55-64	18.1	18.3	20.0	19.2	15.0
65-74	8.6	7.7	8.4	5.1	12.8
75 and above	0.8	0.5	0.9	0.5	1.3
<i>Education (%)</i>					
Less than High School	1.9	2.9	0.9	1.9	2.2
High School Grad	20.0	16.8	22.7	20.6	19.9
Some College or Assoc. Degree	36.3	37.0	35.6	34.6	38.1
Bachelor's Degree	28.9	30.8	28.9	29.0	27.0
Graduate of Professional	12.8	12.5	12.0	14.0	12.8
<i>Income (%)</i>					
Less than \$10,000	5.2	5.8	5.3	6.5	3.1
\$10,000 to 29,999	20.6	20.7	20.9	19.6	21.2
\$30,000 to 49,999	21.8	20.2	21.8	20.6	24.3
\$50,000 to 69,999	20.0	21.6	19.1	17.8	21.7
\$70,000 to 89,999	12.4	14.4	14.2	8.9	11.9
\$90,000 to 149,999	15.6	14.9	12.9	20.6	14.2
\$150,000 and more	4.5	2.4	5.8	6.1	3.5

Table B. 3: Question items for the measure of trust in a charity

How strongly do you agree or disagree with the following statements about [CHARITY NAME] ?

Trust (α of 0.94)	
1.	I would trust this nonprofit to always act in the best interest of the cause
2.	I would trust this nonprofit to conduct their operations ethically
3.	I would trust this nonprofit to use donated funds appropriately
4.	I would trust this nonprofit not to exploit their donors
5.	I would trust this nonprofit to use fundraising techniques that are appropriate and sensitive

Table B. 4: Question items for the measure of dispositional (personality) trust

How strongly do you agree or disagree with the following as statements that apply to you?

Dispositional (Personality) Trust ($\alpha = 0.83$)

1. I believe that others have good intentions
 2. I trust what people say
 3. I believe that people are basically moral
 4. I believe in human goodness
 5. I think that all will be well
 6. I suspect hidden motives in others (Reverse coded)
 7. I am wary of others (Reverse coded)
 8. I believe that people are essentially evil (Reverse coded)
-

Table B. 5: Question items for the measure of altruism

How strongly do you agree or disagree with the following as statements that apply to you?

Altruism ($\alpha = 0.79$)

1. I anticipate the needs of others
 2. I love to help others
 3. I am concerned about others
 4. I have a good word for everyone
 5. I am indifferent to the feelings of others (Reverse coded)
 6. I make people feel uncomfortable (Reverse coded)
 7. I turn my back on others (Reverse coded)
 8. I take no time for others (Reverse coded)
-

Table B. 6: Question items for the measure of trust in nonprofits in general

How strongly do you agree or disagree with the following statements about nonprofits in general?

Trust in nonprofits in general ($\alpha = 0.86$)

1. Generally, nonprofits operate effectively. [competence]
 2. Nonprofits in general are capable in carrying out their missions. [competence]
 3. Nonprofits in general care about citizens' well-being. [benevolence]
 4. In general, nonprofits honor their commitments. [honesty]
-

Table B. 7: Question items for the measure of perceived performance

How strongly do you agree or disagree with the following statements about [CHARITY NAME]?

Perceived Performance of the Organization ($\alpha = 0.68 - 0.73$)

1. This nonprofit is most likely to have an impact on this cause
2. This nonprofit efficiently spends money on this cause

Table B. 8: Question items for the measure of emotional utility

Emotional Utility ($\alpha = 0.86 - 0.91$)

1. I give to this nonprofit because I would feel guilty if I didn't
2. If I didn't give to this nonprofit, I would feel bad about myself

Table B. 9: Question items for the measure of familial utility

Table C-7.

Familial Utility ($\alpha = 0.75 - 0.76$)

1. I give money to this nonprofit in memory of a loved one
2. I felt that someone I know might benefit from my support
3. My family had a strong link to this nonprofit

APPENDIX C

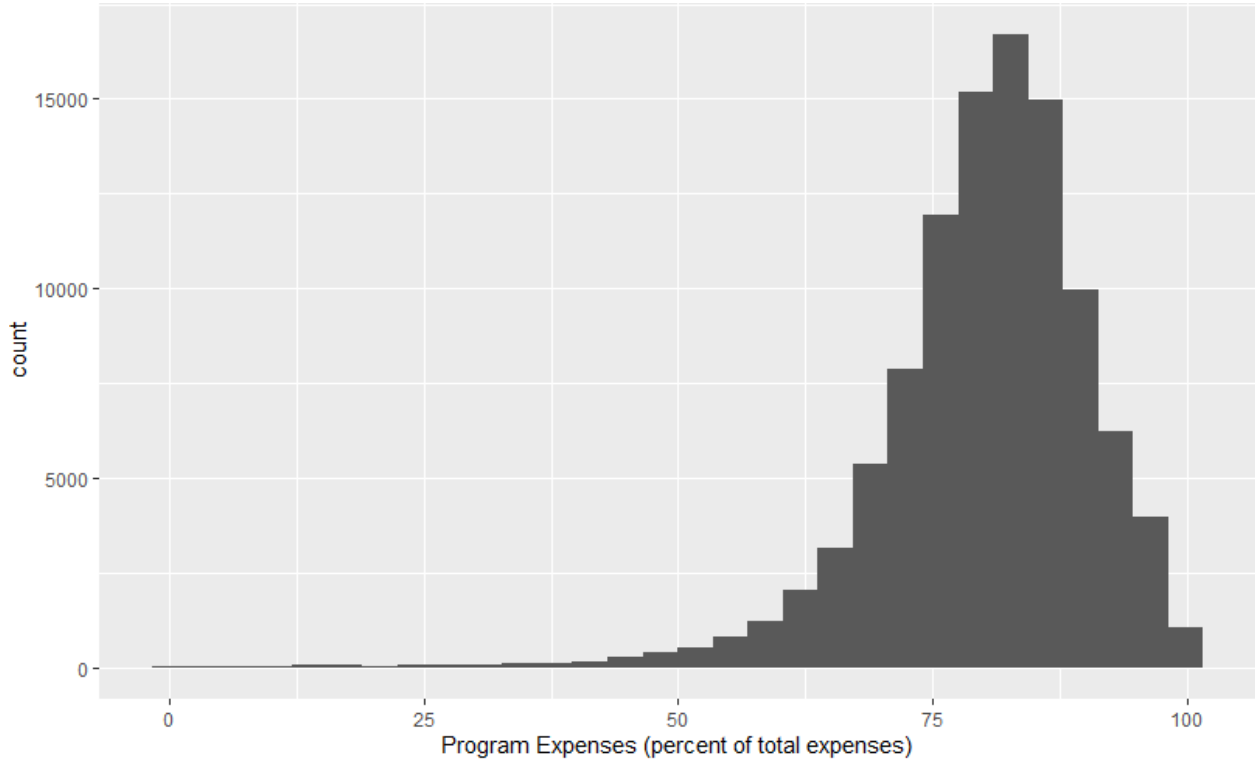


Figure C. 1: Charity program expenses for 8640 charities rated during 2000-2018 (percent)

Table C. 1: Charity program expenses for 8640 charities rated during 2000-2018 (percent)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
0.10	75.10	81.30	80.07	86.70	100.00	83

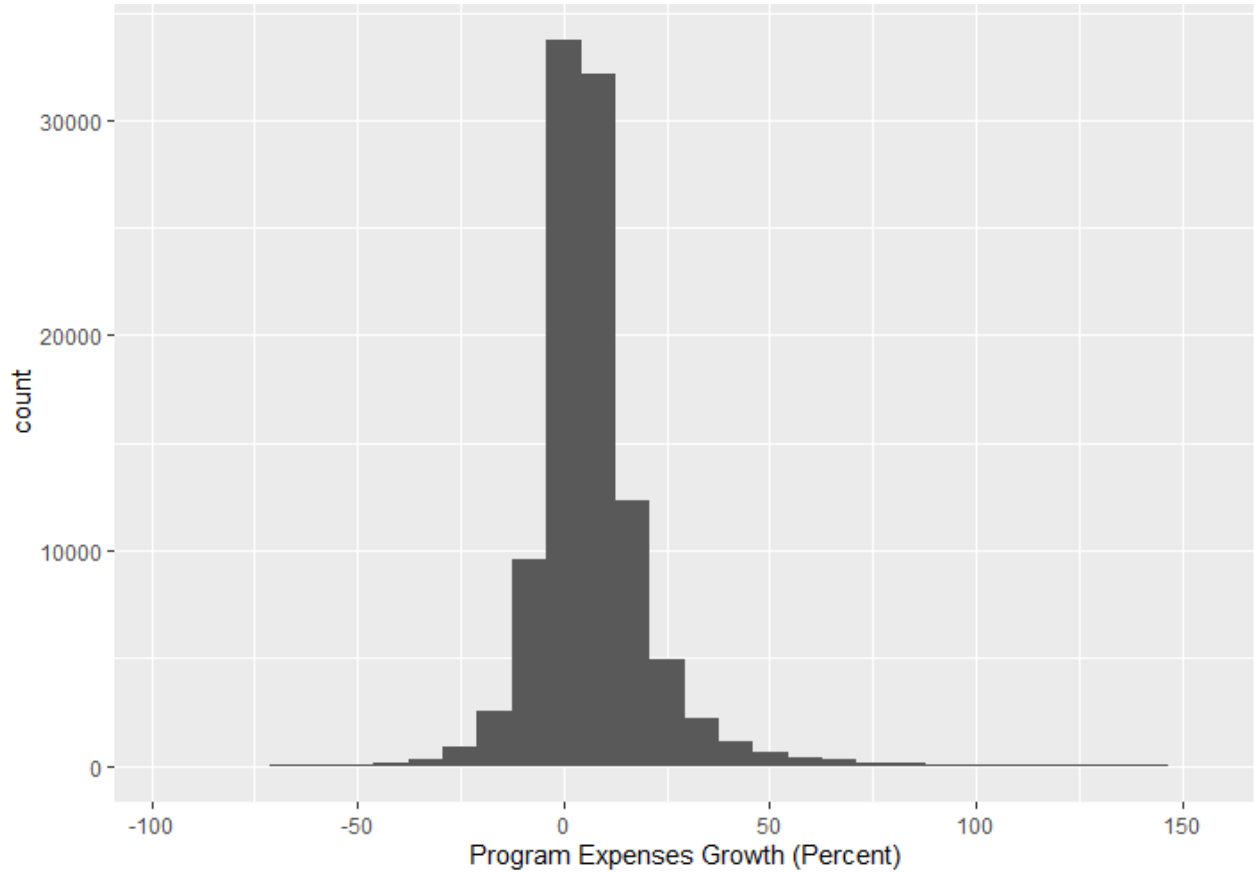


Figure C. 2: Charity program expenses average annual growth for 8640 charities rated during 2000-2018 (percent)

Table C. 2: Charity program expenses average annual growth for 8640 charities rated during 2000-2018 (percent)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
-93.40	-0.20	4.90	6.97	11.40	1007.40	792

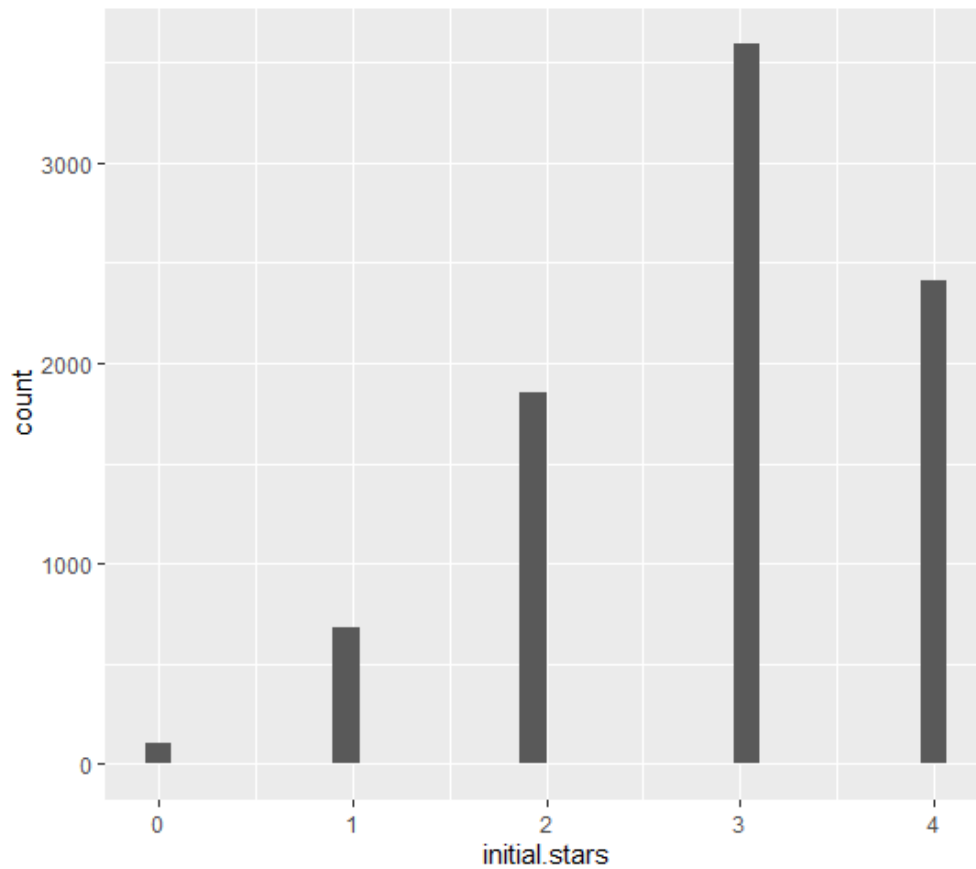


Figure C. 3: Distribution of initial overall ratings for 8640 charities rated during 2000-2018 (measured in stars)

Table C. 3: Distribution of initial overall ratings for 8640 charities rated during 2000-2018 (measured in stars)

	0	1	2	3	4
n	102	682	1855	3588	2413
%	1.18	7.89	21.47	41.53	27.93

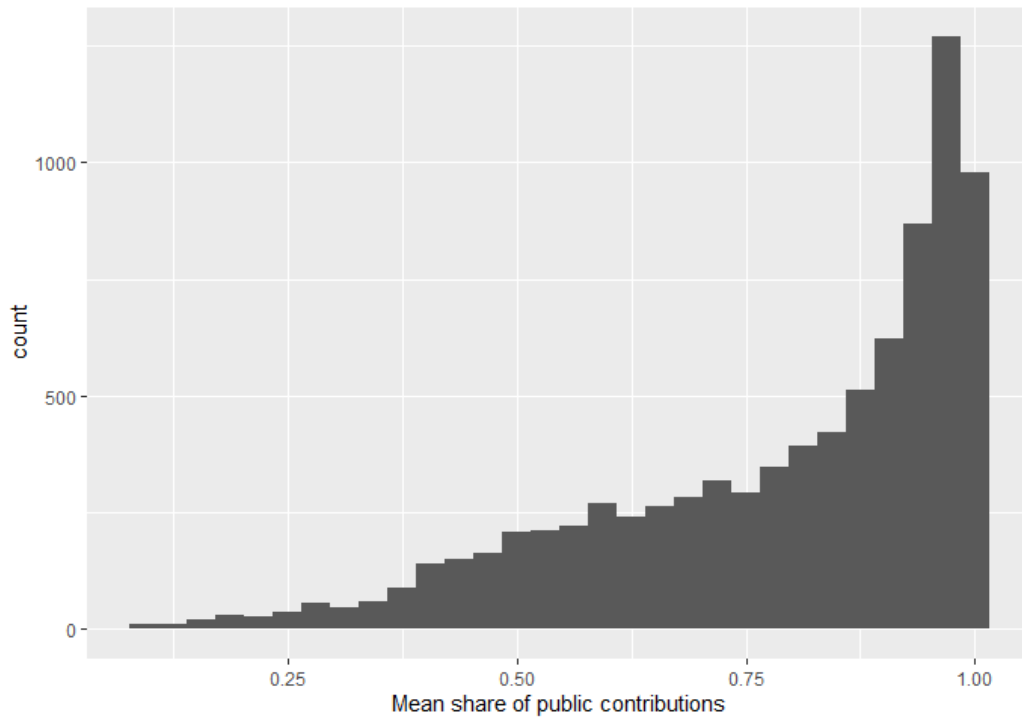


Figure C. 4: Distribution of the mean share of public contributions in nonprofit agency income portfolios for 8640 charities rated during 2000-2018.

Table C. 4: Descriptive statistics for the distribution of the mean share of public contributions in nonprofit agency income portfolios for 8640 charities rated during 2000-2018.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.62	0.86	0.77	0.96	1.0

BIBLIOGRAPHY

- Akerlof, G. A. (1970). The Market for “Lemons”: Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3), 488-500. doi:10.2307/1879431
- Attkisson, S. (2009). Student Loan Charity Under Fire. Retrieved from <http://www.cbsnews.com/news/student-loan-charity-under-fire/>
- Bækgaard, M., & Serritzlew, S. (2015). Interpreting performance information: Motivated reasoning or unbiased comprehension. *Public Administration Review*, 76.
- Barnow, B. S., & Heinrich, C. J. (2010). One standard fits all? The pros and cons of performance standard adjustments. *Public Administration Review*, 70(1), 60-71.
- BBB Wise Giving Alliance. (2013). Audited Financial Statements and Supplementary Information Retrieved from <http://www.give.org/globalassets/wga/annual-reports/bbb-wga-2013-audited-financial-statements.pdf>
- BBB Wise Giving Alliance. (2019). *2018 Annual Report*. Retrieved from <https://www.give.org/docs/default-source/default-document-library/bbb-wga-annual-report-2018.pdf>
- BBB Wise Giving Alliance. (2020). More About Us. Retrieved from <https://give.org/about-bbb-wga/more-about-us/>
- Behn, R. D. (2003). Why measure performance? Different purposes require different measures. *Public Administration Review*, 63(5), 586-606.
- Bennett, R., & Savani, S. (2003). Predicting the accuracy of public perceptions of charity performance. *Journal of Targeting, Measurement and Analysis for Marketing*, 11(4), 326-342.
- Better Business Bureau. (2020). BBB Accreditation for Charities. Retrieved from <https://www.bbb.org/cincinnati/charities-donors/cincinnati-bbb-foundations-charity-education-services/>
- Brody, E. (2002). Accountability and Public Trust. In L. M. Salamon (Ed.), *The state of nonprofit America* (1 ed.): Brookings Institution Press.
- Brown, A. L., Meer, J., & Williams, J. F. (2014). *Social Distance and Quality Ratings in Charity Choice*. Retrieved from

- Brown, E., & Martin, D. (2011). Individual Giving and Volunteering. In L. M. Salamon (Ed.), *The state of nonprofit America* (2 ed., pp. 495-517): Brookings Institution Press.
- Brown, E., & Slivinski, A. (2006). Nonprofit organizations and the market. *The nonprofit sector: A research handbook*, 2, 140-158.
- Burt, C. D., & Dunham, A. H. (2009). Trust generated by aid agency web page design. *International Journal of Nonprofit and Voluntary Sector Marketing*, 14(2), 125-136.
- Burt, C. D., & Gibbons, S. (2011). The effects of donation button design on aid agency transactional trust. *International Journal of Nonprofit and Voluntary Sector Marketing*, 16(2), 183-194. doi:10.1002/nvsm.412
- Butera, L., & Horn, J. R. (2014). Good news, bad news, and social image: The market for charitable giving. *George Mason University Interdisciplinary Center for Economic Science (ICES) Working Paper*.
- Capeci, J. (1991). Credit risk, credit ratings, and municipal bond yields: a panel study. *National Tax Journal*, 41-56.
- Carnochan, S., Samples, M., Myers, M., & Austin, M. J. (2014). Performance measurement challenges in nonprofit human service organizations. *Nonprofit and Voluntary Sector Quarterly*, 43(6), 1014-1032.
- Charity Navigator. (2013). Annual Report 2013. Retrieved from http://www.charitynavigator.org/docs/CN_2013_annual_report.pdf
- Charity Navigator. (2015). Mission. Retrieved from <http://www.charitynavigator.org/index.cfm?bay=content.view&cpid=17#.VXuKJflVikq>
- Charity Navigator. (2016a). Financial Score Conversions and Tables. Retrieved from <https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=48#PerformanceMetricSeven>
- Charity Navigator. (2016b). Rating System Evolution. Retrieved from <https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=2200>
- Charity Navigator. (2020a). Charity Navigator's Methodology. Retrieved from <https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=5593#rating>
- Charity Navigator. (2020b). How Do We Rate Charities' Accountability and Transparency? Retrieved from <https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=1093>

- Charity Navigator. (2020c). What Criteria Must A Charity Meet To Be Rated? Retrieved from <https://www.charitynavigator.org/index.cfm?bay=content.view&cpid=32>
- Charity Watch. (2012). "Not So" GreatNonprofits. Retrieved from <https://www.charitywatch.org/charitywatch-articles/-34-not-so-34-greatnonprofits/16>
- Charity Watch. (2018). CharityWatch Hall of Shame: The Personalities Behind Charity Scandals. Retrieved from <https://www.charitywatch.org/charity-donating-articles/charitywatch-hall-of-shame>
- Charity Watch. (2020). Mission Statement. Retrieved from <https://www.charitywatch.org/about-charitywatch/mission-goals>
- CharityWatch. (2020a). CharityWatch Difference. Retrieved from <https://www.charitywatch.org/about-charitywatch/charitywatch-difference>
- CharityWatch. (2020b). Frequently Asked Questions. Retrieved from https://www.charitywatch.org/about-charitywatch/faq#charity_selection
- CharityWatch. (2020c). Our Charity Rating Process. Retrieved from <https://www.charitywatch.org/our-charity-rating-process>
- Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1), 1-8. doi:<https://doi.org/10.1016/j.jebo.2011.08.009>
- Chatterji, A. K., & Toffel, M. W. (2010). How firms respond to being rated. *Strategic Management Journal*, 31(9), 917-945.
- Chhaochharia, V., & Ghosh, S. (2008). *Do Charity Ratings Matter?* Retrieved from <https://EconPapers.repec.org/RePEc:fal:wpaper:08001>
- Cyert, R. M., & March, J. G. (1963). A behavioral theory of the firm. *Englewood Cliffs, NJ*, 2(4), 169-187.
- DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, 48(2), 147-160. doi:10.2307/2095101
- Direct Relief. (2020). Efficiency & Effectiveness. Retrieved from https://www.directrelief.org/about/charity-rating/?gclid=CjwKCAjw_LL2BRakEiwAv2Y3Sf4GJnV1x2XD71pq73mUttfkKVJjp5u3yMJD3PVABKG4FKgstaNLABoCx2kQAvD_BwE

- Ebrahim, A., & Rangan, V. K. (2010). *The limits of nonprofit impact: A contingency framework for measuring social performance*. Retrieved from
- Elsbach, K. D., & Kramer, R. M. (1996). Members' responses to organizational identity threats: Encountering and countering the Business Week rankings. *Administrative Science Quarterly*, 442-476.
- Eng, M. (2011). Watchdogging the charity watchdogs. *The Seattle Times*. Retrieved from <https://www.seattletimes.com/life/lifestyle/watchdogging-the-charity-watchdogs/>
- Environmental Defense Fund. (2020). Our charity ratings. Retrieved from <https://www.edf.org/our-charity-ratings>
- Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds1. *American journal of sociology*, 113(1), 1-40.
- Figlio, D. N., & Kenny, L. W. (2009). Public sector performance measurement and stakeholder support. *Journal of Public Economics*, 93(9), 1069-1077.
- Gilkeson, N. (2006). For-profit scandal in the nonprofit world: Should states force Sarbanes-Oxley provisions onto nonprofit corporations. *Geo. LJ*, 95, 831.
- Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding overhead aversion in charity. *Science*, 346(6209), 632-635.
- Goldberg, E. (2015). Watchdogs Call For More Oversight Of Red Cross Amid Donation Scandal. Retrieved from http://www.huffingtonpost.com/entry/watchdogs-call-for-more-oversight-of-red-cross-amid-donation-scandal_55f9c8e6e4b0fde8b0ccac98
- Google. (2013). MISSION 501(c)(3): Driving Donations, Digitally. Retrieved from <https://www.thinkwithgoogle.com/research-studies/digital-non-profits-study.html>
- Gordon, T. P., Knock, C. L., & Neely, D. G. (2009). The role of rating agencies in the market for charitable contributions: An empirical test. *Journal of Accounting and Public Policy*, 28(6), 469-484.
- Gormley, W. T. (2003). *Using organizational report cards*. Paper presented at the National Public Management Research Conference, October.
- Gormley, W. T., & Weimer, D. L. (1999). *Organizational report cards*: Harvard University Press.
- Graham, M. (2000). Regulation by shaming. *Atlantic Monthly*, 285(4), 36-40.

- GreatNonprofits. (2015). About GreatNonprofits. Retrieved from <http://www.about.greatnonprofits.org/#!about/cm8a>
- Grimmelikhuijsen, S. G., & Meijer, A. J. (2012). The effects of transparency on the perceived trustworthiness of a government organization: Evidence from an online experiment. *Journal of Public Administration Research and Theory*, mus048.
- GuideStar. (2020). GuideStar Seals of Transparency. Retrieved from <https://learn.guidestar.org/seals>
- GuidStar. (2020). 2020 GuideStar Profile Standard. Retrieved from <https://learn.guidestar.org/hubfs/Docs/2020-GuideStar-Profile-Standard.pdf>
- Hansmann, H. B. (1980). The Role of Nonprofit Enterprise. *The Yale Law Journal*, 89(5), 835-901. doi:10.2307/796089
- Hart, T., Greenfield, J. M., & Haji, S. D. (Eds.). (2007). *People to People Fundraising: Social Networking and Web 2.0 for Charities*, by T. Hart, J. M. Greenfield, and S. D. Haji: Hoboken, N.J. : Wiley, c2007.
- Herzlinger, R. E. (1995). Can public trust in nonprofits and governments be restored? *Harvard Business Review*, 74(2), 97-107.
- Hirschman, A. O. (1970). *Exit, voice, and loyalty : responses to decline in firms, organizations, and states / Albert O. Hirschman*: Cambridge, Massachusetts : Harvard University Press.
- Hoffman, S. (2006). For U.S. Charities, a Crisis of Trust. Scandals, Accountability Problems Combine to Undermine Public Support. Retrieved from http://www.nbcnews.com/id/15753760/ns/us_newsgiving/t/uscharitiescrisistrust/#.VkqXpLerTDe
- Hunter, T., & Bansal, P. (2007). How standard is standardized MNC global environmental communication? *Journal of Business Ethics*, 71(2), 135-147.
- ImpactMatters. (2017). *Form 990 for the 2017 calendar year. Return of organization Exempt from Income Tax*. Retrieved from https://s3.amazonaws.com/impactmatters-production/pdfs/Form_990_Public_Copy_12-31-17.pdf
- ImpactMatters. (2020a). Frequently Asked Questions (for donors). Retrieved from <https://www.impactmatters.org/about/faq/>
- ImpactMatters. (2020b). Frequently Asked Questions (for nonprofits). Retrieved from <https://www.impactmatters.org/nonprofit-center/faq/>

- ImpactMatters. (2020c). Impact Rating Standard. Retrieved from <https://www.impactmatters.org/methodology/impact-rating-standard/impact-rating-standard.html>
- Interactive, H. (2006). While a third of adults think the nonprofit sector in the United States is headed in the wrong direction, a vast majority of households have donated to charities in the past year. The Harris Poll (No. 33). Retrieved July 9, 2006. In.
- James, O. (2011). Performance measures and democracy: Information effects on citizens in field and laboratory experiments. *Journal of Public Administration Research and Theory*, 21(3), 399-418.
- Jin, G. Z., & Leslie, P. (2003). The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards. *The Quarterly Journal of Economics*, 118(2), 409-451. doi:10.2307/25053911
- Jin, G. Z., & Leslie, P. (2009). Reputational incentives for restaurant hygiene. *American Economic Journal: Microeconomics*, 1(1), 237-267.
- Jin, G. Z., & Whalley, A. (2007a). The Power of Attention: Do Rankings Affect the Financial Resources of Public Colleges? *NBER Working Paper Series*, 12941.
- Jin, G. Z., & Whalley, A. (2007b). The Power of Information: How Do US News Rankings Affect the Financial Resources of Public Colleges?
- Johnson, C. L., & Kriz, K. A. (2002). Impact of three credit ratings on interest cost of state GO bonds. *Municipal Finance Journal*, 23(1), 1-16.
- Kelly, K. S. (1998). *Effective fund-raising management Kathleen S. Kelly*. Mahwah, N.J.: Mahwah, N.J. : Lawrence Erlbaum Associates.
- Lampkin, L. M., Winkler, M. K., Kerlin, J., Hatry, H. P., Natenshon, D., Saul, J., . . . Seshadri, A. (2007). *Building a common outcome framework to measure nonprofit performance*. Retrieved from <http://webarchive.urban.org/publications/411404.html>
- Lewis, B. W. (2014). The Paradox of Recognizing Responsibility: Why Positive Social Ratings Can Lead to Reductions in Corporate Social Performance. Available at SSRN 2390074.
- Light, P. C. (2008). *How Americans view charities: A report on charitable confidence, 2008*: Brookings Institution.
- Lizzeri, A. (1999). Information revelation and certification intermediaries. *The RAND Journal of Economics*, 214-231.

- Longo, D. R., Land, G., Schramm, W., Fraas, J., Hoskins, B., & Howell, V. (1997). Consumer reports in health care: Do they make a difference in patient care? *JAMA*, 278(19), 1579-1584. doi:10.1001/jama.1997.03550190043042
- Lowell, S., Trelstad, B., & Meehan, B. (2005). The ratings game. *Stanford Social Innovation Review*, 3, 38-45.
- Luca, M. (2011). Reviews, reputation, and revenue: The case of Yelp. com. *Com* (September 16, 2011). *Harvard Business School NOM Unit Working Paper*(12-016).
- Lynch-Cerullo, K., & Cooney, K. (2011). Moving from outputs to outcomes: A review of the evolution of performance measurement in the human service nonprofit sector. *Administration in Social Work*, 35(4), 364-388.
- MacLaughlin, S. (2015). *Charitable Giving Report. How Nonprofit Fundraising Performed in 2014*. Retrieved from www.blackbaud.com: www.blackbaud.com
- Martins, L. L. (2005). A model of the effects of reputational rankings on organizational change. *Organization Science*, 16(6), 701-720.
- Moe, T. M. (1984). The New Economics of Organization. *American Journal of Political Science*, 28(4), 739-777. Retrieved from <http://www.jstor.org/stable/2110997>
- Moxham, C. (2009). Performance measurement: Examining the applicability of the existing body of knowledge to nonprofit organisations. *International Journal of Operations & Production Management*, 29(7), 740-763.
- National Council of Nonprofit Associations and the National Human Services Assembly. (2005). Rating the Raters: An Assessment of Organizations and Publications That Rate/Rank Charitable Nonprofit Organizations. Retrieved from <http://www.nationalassembly.org/uploads/publications/documents/ratingtheraters.pdf>
- NPTrust. (2015). Charitable Giving Statistics. Retrieved from <http://www.nptrust.org/philanthropicresources/charitablegivingstatistics/>
- O'Donnell, J. (2012). BBB's charity ratings, seal of approval under fire. *USA Today*, 2015(7/10). Retrieved from <http://www.usatoday.com/story/money/personalfinance/2012/12/27/better-business-bureau-charity-ratings-donations/1636957/>
- Peng, S., Kim, M., & Deat, F. (2019). The Effects of Nonprofit Reputation on Charitable Giving: A Survey Experiment. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 30(4), 811-827. doi:10.1007/s11266-019-00130-7

- Peterson, E. D., DeLong, E. R., Jollis, J. G., Muhlbaier, L. H., & Mark, D. B. (1998). The effects of New York's bypass surgery provider profiling on access to care and patient outcomes in the elderly. *Journal of the American College of Cardiology*, 32(4), 993-999.
- Poister, T. H., Aristigueta, M. P., & Hall, J. L. (2014). *Managing and Measuring Performance in Public and Nonprofit Organizations: An Integrated Approach*: John Wiley & Sons.
- Pope, D. G. (2009). Reacting to rankings: evidence from "America's Best Hospitals". *Journal of health economics*, 28(6), 1154-1165.
- Powell, W. W., & Steinberg, R. (2006). *The nonprofit sector: a research handbook* (2nd ed. ed.). Yale University Press: New Haven.
- Rhode, D. L., & Packel, A. K. (2009). Ethics and nonprofits. *Stanford Social Innovation Review*, 7(3), 28-35.
- Rovner, M., Loeb, P., McCarthy, D., & Johnston, M. (2013). The Next Generation of American Giving: The Charitable Habits of Generations Y, X, Baby Boomers, and Matures. *Blackbaud* (August 2013). <http://npengage.uberflip.com/i/147711> (accessed May 5, 2014).
- Rowe, G. (2012). Performance measurement. *Introduction to Nonprofit Management: Text and Cases*, 129.
- Salamon, L. M. (2002). *The state of nonprofit America*. Brookings Institution Press :: Washington, D.C.
- Salamon, L. M. (2012). *The state of nonprofit America*: Brookings Institution Press.
- Sargeant, A., Ford, J. B., & West, D. C. (2006). Perceptual determinants of nonprofit giving behavior. *Journal of Business Research*, 59(2), 155-165.
- Sauder, M., & Espeland, W. N. (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, 74(1), 63-82.
- Saxton, G. D., & Guo, C. (2011). Accountability online: Understanding the web-based accountability practices of nonprofit organizations. *Nonprofit and Voluntary Sector Quarterly*, 40(2), 270-295.
- Scanlon, D. P., Chernew, M., Sheffler, S., & Fendrick, A. (1998). Health plan report cards: exploring differences in plan ratings. *The Joint Commission journal on quality improvement*, 24(1), 5-20.

- Sharkey, A. J., & Bromley, P. (2014). Can ratings have indirect effects? Evidence from the organizational response to peers' environmental ratings. *American Sociological Review*, 0003122414559043.
- Silvergleid, J. E. (2003). Effects of watchdog organizations on the social capital market. *New Directions for Philanthropic Fundraising*, 2003(41), 7-26. doi:10.1002/pf.38
- Singh, J. V. (1986). PERFORMANCE, SLACK, AND RISK TAKING IN ORGANIZATIONAL DECISION MAKING. *Academy of Management Journal*, 29(3), 562-585. doi:10.2307/256224
- Sloan, M. F. (2009). The effects of nonprofit accountability ratings on donor behavior. *Nonprofit and Voluntary Sector Quarterly*, 38(2), 220-236.
- Steinberg, R. (2006). Economic Theories of Nonprofit Organizations In *The nonprofit sector: A research handbook* (2 ed., pp. 117-139).
- Szper, R., & Prakash, A. (2011). Charity Watchdogs and the Limits of Information-Based Regulation. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 22(1), 112-141. Retrieved from <http://www.jstor.org/stable/27928253>
- The BBB Wise Giving Alliance. (2015). More About Us. Retrieved from <http://www.give.org/about-bbb-wga/more-about-us/?id=239073>
- The Washington Post. (2018). The Oxfam scandal shows that, yes, nonprofits can behave badly. So why aren't they overseen like for-profits? Retrieved from <https://www.washingtonpost.com/news/monkey-cage/wp/2018/02/19/the-oxfam-scandal-shows-that-yes-nonprofits-can-behave-badly-so-why-arent-they-overseen-like-for-profits/>
- Tremblay-Boire, J., & Prakash, A. (2017). Will You Trust Me?: How Individual American Donors Respond to Informational Signals Regarding Local and Global Humanitarian Charities. *Voluntas: International Journal of Voluntary and Nonprofit Organizations*, 28(2), 621-647. doi:10.1007/s11266-016-9782-4
- Trussel, J. M., & Parsons, L. M. (2007). Financial reporting factors affecting donations to charitable organizations. *Advances in Accounting*, 23, 263-285.
- Vesterlund, L. (2006). Why do people give. *The nonprofit sector: A research handbook*, 2, 168-190.
- Wholey, J. S., & Hatry, H. P. (1992). The Case for Performance Monitoring. *Public Administration Review*, 52(6), 604-610. doi:10.2307/977173

Young, D. R. (2013). *If not for profit, for what?* : (1983 Print Edition) Lexington Books.

Zhe Jin, G., Kato, A., & List, J. A. (2010). That's news to me! Information revelation in professional certification markets. *Economic Inquiry*, 48(1), 104-122.

VITA

Iurii (Yuriy) Davydenko was born in Lviv, Ukraine. He received a Specialist degree in Acoustical Engineering from Kyiv Polytechnic Institute in 2000, and a Specialist degree in Management from Kyiv National Economic University in 2006. During that time, his professional work focused on Ukraine's social and economic reforms to transition to a market-based, democratic society.

In 2008, he received an Edmund S. Muskie Graduate Fellowship from the United States Department of State and moved to Atlanta, GA to begin studies in the field of public policy and administration. He earned a Master's Degree in Public Administration from Georgia State University in 2010 before starting his doctorate in Public Policy. His research aims at understanding how various policies, programs, tools, and institutions affect individual and organizational performance and outcomes. In particular, he is interested in information-based tools, including performance management. Outside of social science research, he enjoys working with technology and developing expertise in data science.

Contact email:

ygdavydenko@gmail.com

DISCLOSURE STATEMENT

Funding for this research was provided through AYS Dissertation Fellowship Grant from the GSU Foundation. The author declares that there are no conflicts of interest that relate to the research, authorship, or publication of this work.