

The Ratings Game

Evaluating the three groups that rate the charities

By Stephanie Lowell, Brian Trelstad, & Bill Meehan

Stanford Social Innovation Review
Summer 2005

Copyright © 2005 by Leland Stanford Jr. University
All Rights Reserved

DO NOT COPY



3 0 20 2 9
always 19 20 23
when 5 2 3 real 4
1 2 ?

3 27 5
yes
8 right
2
sometimes

230 50 then 3 2
possibly
7
1
out
9
49
15
9
if
down
5
362
2
never
250 4
0
between

by STEPHANIE LOWELL, BRIAN TRELSTAD, & BILL MEEHAN

the ratings game

The tsunami that struck South Asia in December 2004 will be remembered not only for the scale of the human misery it caused, killing hundreds of thousands and displacing millions, but also for the unprecedented global outpouring of charity it evoked. Within a few weeks of the disaster, over \$400 million (on the way to an estimated total of \$1 billion) had been raised by U.S. aid organizations alone; furthermore, a large proportion of those donations was made via the Internet.

“The response has been unprecedented,” says Mike Kiernan of Save the Children USA, “greater than any other disaster or crisis in (our) more than 70 years of operation.” By April, roughly 20 percent of the \$63 million Save the Children USA had collected for tsunami victims had come in through its Web site – a 100-fold increase from pre-tsunami levels. Other groups reported a similar shift in giving patterns.

In response, some of the charities benefiting from this surge in donations started behaving in new ways, too. For example, the U.S. branch of Doctors Without Borders announced a week after the disaster that it had already raised as much money as it could responsibly use, given the lim-

ited scale of its operations in the affected areas. Another group, Direct Relief International, assured donors it was depositing its flood of donations into a separate bank account, and that the salaries of its employees would not be paid out of these donations, as part of its effort to maximize the amount that would reach the victims.

If all of this heralds a new age in philanthropy, where the Internet will be a dominant force in charity, bringing a new sense of accountability and transparency to the process, there are three online services already in place that stand to benefit. These three have built Web-based charity rating services, available to give prospective donors information and guidance about the groups they wish to consider for support. They are the BBB Wise Giving Alliance, which uses a set of 20 standards to monitor the operations and financial stability of national charities, and uses a “pass-fail” system of grading; Charity Navigator, which rates non-

*Evaluating
the
three groups
that rate
the
charities*

ILLUSTRATION BY JOYCE HESSELBERTH/IMAGES.COM

profits on a set of organizational efficiency and organizational capacity metrics with a “star-based” system of one to four stars, using algorithms derived from publicly reported financial data; and the American Institute of Philanthropy (AIP), which rates nonprofits with grades of “A+” through “F” using financial ratios and analysis of charities’ financial statements, including their 990s and audited financials.”¹

Not surprisingly, all three of these services saw their user traffic grow exponentially in the wake of the tsunami. At Charity Navigator, for example, traffic grew tenfold, from an average of 5,000 unique visitors a day to over 50,000 during the week following the tragedy.

Over the past few years, each of these ratings sites has sought to establish itself as *the* authority for donors seeking information to guide their giving decisions. And their influence is starting to be felt, as many nonprofits now proudly tout their high ratings from these organizations on their Web sites (“Save the Children awarded 4-star rating from Charity Navigator”), and portals such as Earthlink direct users to the “top-rated charities” identified by the ratings agencies.

We conducted a detailed study of the agencies to determine how useful a service they provide. The results were sobering: Our review of their methodologies indicates that these sites individually and collectively fall well short of providing meaningful guidance for donors who want to support more efficient and effective nonprofits.

Based on our study, the major weaknesses of the ratings agencies are threefold: They rely too heavily on simple analysis and ratios derived from poor-quality financial data; they overemphasize financial efficiency while ignoring the question of pro-



Sri Lanka's eastern coast was devastated by last December's tsunami. The disaster led to an unprecedented outpouring of charity all over the world.

gram effectiveness; and they generally do a poor job of conducting analysis in important qualitative areas such as management strength, governance quality, or organizational transparency.

To be fair, these are early days for the ratings business; all of the sites are less than six years old² and each is still working on improving its methodology, growing its user base, and developing a sustainable business model for its services.

But as traffic to these rating sites grows, and donors make important decisions using potentially misleading data and analysis, the agencies' potential to do harm may outweigh their ability to inform. In this article, we review some of the strengths and weaknesses of the ratings agencies and consider how to build a more effective and transparent system for nonprofit ratings and evaluation.

Rating the Raters

To assess how well the ratings agencies do their job, we put them to a two-part test. First, we put ourselves in the position of a hypothetical donor for tsunami relief. How helpful would these agencies have been? Second, we conducted a more thorough review of each of the rating agencies' services, tried to understand their evaluation methodology, and interviewed their leaders in an effort to understand their potential for guiding effective donor decision making.

First, we tested how each rating agency would rank seven of the 10 largest recipients of tsunami aid (see chart, p. 41).³ All seven got a “pass” from Wise Giving, three or four stars from

STEPHANIE LOWELL is an independent nonprofit consultant and an alumnus of McKinsey & Company's Boston office. She can be reached at stephanie@smlassociates.org.

BRIAN TRELSTAD is the CFO of Acumen Fund and an alumnus of McKinsey's New Jersey office. He can be reached at btrelstad@acumen-fund.org.

WILLIAM F. MEEHAN III is a director of McKinsey & Company, senior lecturer in strategic management and Class of 1978 Lecturer (2004-2005) at Stanford University Graduate School of Business. He is also the chairman of Philanthropic Research Inc., the parent of GuideStar. He can be reached at meehan_william@gsb.stanford.edu.

PHOTOGRAPH BY BRUNO STEVENS / COSMOS AURORA

RATINGS FOR TOP RECIPIENTS OF TSUNAMI AID IN THE UNITED STATES

	Wise Giving Alliance	Charity Watch	Charity Navigator	Forbes*
American Red Cross (\$236m)	Meets Standards	A-	4 stars (64.11)	1) 91% 2) 80% 3) 166%
US Fund for UNICEF (\$68m)	Meets Standards	C-**	4 stars (61.29)	1) 89% 2) 91% 3) 103%
Save the Children USA (\$41m)	Meets Standards	A	4 stars (61.59)	1) 91% 2) 87% 3) 92%
World Vision USA (\$29m)	Meets Standards	B+	4 stars (61.02)	1) 81% 2) 87% 3) 92%
CARE USA (\$25m)	Meets Standards	A-	4 stars (66.73)	1) 92% 2) 89% 3) 85%
Doctors Without Borders (\$20m)	Meets Standards	A	3 stars (54.80)	1) 86% 2) 89% 3) 92%
AmeriCares (\$20m)	Meets Standards	A	4 stars (62.40)	1) 99% 2) 99% 3) 94%

*Forbes annually ranks charities using three ratios: 1) Charitable services as percent of total expenses; 2) Percent of private support remaining after fundraising expenses; and 3) Percent of private support remaining after surplus.

**Charity Watch discounts in-kind revenues in its calculations.

SOURCE: WEB SITES

Charity Navigator, and with the exception of the US Fund for UNICEF, a B+ or higher from AIP. With the exception of the US Fund for UNICEF, whose poor grade isn't explained on the AIP Web site, does the analysis really help you make an affirmative decision? Is there enough information to distinguish one of these charities from the rest? While it is certainly reassuring to know that contributing to any one of these seems to be a reasonable choice (with the debatable exception of UNICEF), the ratings sites collectively fail the test to actually inform positive donor choice about allocating scarce capital among competing options.

For our more in-depth review, we analyzed each agency's finances and their Web sites. We then interviewed the senior leadership at each organization to understand their methodology. Finally, we created a qualitative review of what criteria we considered important: the number of nonprofits rated, the range of sources used, and level of interpretation in generating scores, the transparency of their analysis, and their inclusion of non-financial criteria (see chart, p. 42).

BBB Wise Giving Alliance, which is affiliated with the Better Business Bureau, is the oldest of the raters and has the most comprehensive approach, using quantitative and qualitative analysis to pass or fail 500 national charities. Their staff mem-

bers review nonprofits in the areas of financial efficiency and stability, governance and oversight, performance measurement, and the quality and accuracy of the organization's fundraising and informational material. "Finances alone are only a piece of the picture, and in fact can give you a 'false positive' on the health of the organization," notes Bennett Weiner, Wise Giving's chief operating officer.

Wise Giving's 20 evaluation standards were developed over three years with input from hundreds of donors, nonprofit leaders, government regulators, and academics. The agency interacts with the nonprofits they review both by contacting them to discuss issues or concerns, and then follows up by posting an implementation guide on its Web site to ensure that the evaluation process is transparent to all. The drawbacks, however, are that the time-intensive analysis has limited the number of organizations they rate (though they aim to grow to 3,000 by 2007), and that the tool is only useful in weeding out unethical, deceptive, or poorly managed organizations, not in helping make distinctions among the majority of nonprofits that "meet standards."

Charity Navigator focuses on helping donors make informed decisions by enhancing the transparency of a range of financial data.⁴ Charity Navigator certainly gets style points for user-

RATINGS AGENCIES VARY ON SEVERAL CHARACTERISTICS

	Number of Charities rated	Diversity of sources in analysis	Amount of manual data analysis	Transparency of methodology	Ease of peer benchmarking	Inclusion of nonfinancial criteria
Wise Giving Alliance	500	H	M	H	L	H
AIP	500	M	M	L	M*	L
Charity Navigator	3,700	L	L	H	H	L
Forbes	200	L	L	H	M	L
GuideStar (analyst reports)	1 million	H	M	H	M	H**

*Challenge of benchmarking is that you can only do so with the print report, not on the Web.

**Organizations can self-report accomplishments, highlights of the past year.

friendliness and visual appeal, bringing together a range of information on 3,700 nonprofits, including financial metrics and a summary of their mission. The site also uses a range of financial ratios and peer benchmarks, taking into account that cost structures (and thus financial metrics) may vary by subsector. For example, food banks, because of their reliance on donated goods, may have less need for a certain level of cash reserves as a percentage of their revenues, or that public broadcasting, because it uses expensive airtime for fundraising, may have slightly higher fundraising costs.

Nonetheless, Charity Navigator's effectiveness is hampered by its exclusive focus on financial analysis derived from only one year of 990 data. "To rely exclusively on data from the 990s is ridiculous," commented Bob Ottenhoff, president of GuideStar, "and it's reckless if a single entry from a single year can materially change a charity's rating. One thing we have learned from looking at millions of 990s as we have scanned them into our database is that they vary tremendously in quality. That some absurdly high proportion of nonprofits report that they spend no money on fundraising is a typical problem with how 990s are filled out." Ottenhoff was referring to a 1999 Urban Institute study that found that 59 percent of 58,000 charities that received public donations either reported zero fundraising expenses or left the fundraising expense line blank on their 990.⁵

Given these widespread concerns about the accuracy and reliability of 990 data, Charity Navigator's ratios, particularly when carried out to the second decimal point, feel a bit arbitrary (see sidebar, p. 43). Furthermore, generating ratios on resource efficiency, even with reliable numbers, only tells you about use of resources, not about the program effectiveness. It's a bit like wine connoisseur Robert Parker giving that Oregon Pinot Noir a 93

not for its taste, but based on the number of grapes used to make the wine.

Trent Stamp, Charity Navigator's executive director, does believe (as a note on the Web site explains) that financials provide only a piece of the full picture of the strength of an organization, but he explains that "ultimately the donor is our customer, and the donor is first and foremost asking for these measures of financial health. You can disagree with the methodology, but it is clear, transparent, and user-friendly" and the relatively automatic financial calculations drawn from public databases of 990s have enabled Charity Navigator to build the largest collection of rated organizations, which they plan to grow to 5,000 rated organizations by the end of this year.

Finally, we come to AIP, which issues letter grades ranging from A+ to F for about 500 nonprofits. On the positive side, it recognizes the limitations of the 990 and thus develops its financial health ratios by analyzing a charity's audited financial statements. AIP's small staff of analysts looks closely at specific calculations, including how nonprofits allocate telemarketing costs, which are often labeled "education and outreach," and in-kind contributions, which they assert are often overvalued, among other practices they think nonprofits use when preparing their 990s to cast a more positive light on their financial position.

While it is true that nonprofits have wide latitude in completing their 990s (and many do go to great lengths to misrepresent their financial information), it is difficult for a donor to understand what specific adjustments AIP made to a given nonprofit's ratings and why. (The printed report shares the adjusted ratio, but not details of the analysis.) Their full report is available by mail⁶ (a curious business practice in the age of the Internet), and provides additional, but still incomplete, insight into the specifics of the analysis on any given organization.

AIP is also not afraid to fail an organization; in fact, they specifically aim to review nonprofits that they feel aren't spending wisely or performing ethically, to help educate the public. "We're really looking at the numbers and what they mean, not just running 990 inputs through an equation," said Daniel Borochoff, AIP's president. "At times we actually find that a nonprofit is selling itself short in the way they report the numbers, and help them fill out the 990 more accurately, but more often we see nonprofits misleading potential donors with the way they report their financials. You have to ask yourself why the other [rating organizations] aren't seeing the same really bad things going on with the numbers at some of these charities," Borochoff commented.

Unfortunately, this "gotcha" mentality and lack of transparency are AIP's biggest shortcomings. A donor sees the score, but only limited explanation, and this approach can cause more harm than good. Ultimately, our analysis led us to the conclusion that none of the three agencies provides sufficient input into donor decision making as a stand-alone source.

Toward a More Effective Rating System

What would it take to build a truly effective ratings system? The existing rating agencies, despite our reservations, have taken a step in the direction toward increasing information transparency and accountability of the sector, but they all still fall short. The limited data they provide can be helpful to the educated donor who uses the information as input into a larger decision-making process, but the uneducated donor is easily misled by some of these oversimplified scores.

Bruce Sievers, former executive director of the Walter and

Elise Haas Fund, says: "Sure these ratings agencies are serving the donor, but it is irresponsible not to educate donors on the many aspects of effectiveness, beyond the financials, even if you can't perfectly measure them all. Many, if not most, important aspects of nonprofit activity are intangible."

A more effective nonprofit rating system should have at least four main components: improved financial data that is reviewed over three to five years and put in the context of narrowly defined peer cohorts; qualitative evaluation of the organization's intangibles in areas like brand, management quality, governance, and transparency; some review of the organization's program effectiveness, including both qualitative critique by objective experts in the field, and, where appropriate, "customer" feedback from either the donor or the aid recipient's perspective; and an opportunity for comment or response by the organization being rated.

First, the financial data needs to be improved, made more reliable, and interpreted in a more sophisticated manner. Efforts are under way at both the federal and state levels to reform nonprofit financial reporting, specifically in setting higher standards for completing 990s and holding leadership more accountable for the numbers. Meanwhile, existing 990 financial data can be analyzed more effectively by looking at three- to five-year time horizons and by comparing the data with narrower peer groups.

What might be needed is an analysis using existing nonprofit financial data to predict some desirable future financial state for a given organization. Perhaps, the nonprofit raters could emulate their for-profit peers (e.g. Moody's), who evaluate the creditworthiness of for-profits, nonprofits, and governments looking to borrow capital. These private credit rating agencies,

The Limitations of Relying on Form 990

While the electronic availability of the IRS Form 990 for most nonprofits is a huge leap forward for the transparency of the nonprofit sector, the unreliability of the information is currently the subject of debate in the United States Senate Finance Committee. The concerns include the lack of nonprofit-specific accounting standards and oversight (comparable to Financial Accounting Standards Board oversight of for-profit accounting) leaving room for interpretation, particularly in areas of valuing in-kind contributions, allocat-

ing administrative expenses, and reporting fundraising expenses. There is also limited board and executive accountability for 990 completion and accuracy, and relatively soft penalties for noncompliance. Finally, extensions and late filings are commonplace, making it increasingly difficult to access timely information about nonprofit financial performance.

Virginia Hodgkinson, a board member of GuideStar, notes that "relying on the 990 as your primary source of information is dangerous. It's a snapshot in time, there are no enforced standards, and it is easy for

each nonprofit to report things differently. For example, some organizations allocate fundraising and overhead costs across categories, others don't." And the Government Accounting Office shares her concern: "Although disclosure of charity spending data can facilitate public oversight, caution in interpreting the data is warranted. No measures are available on the accuracy of the expense data and substantial discretion in allocating the expenses makes use of the data problematic in comparing charities."



// It is irresponsible not to educate donors on the many aspects of effectiveness beyond the financials. //

despite recent criticism for missing the collapses of Enron and Worldcom, can point to a long-term record of success: Fewer than 5 percent of companies rated AAA/AA/A have gone bankrupt over the past 15 years.

Or perhaps, someone could tap into the distributed intelligence of the nonprofit information markets, much like the Iowa Electronic Markets (IEM) have transformed presidential punditry in the United States by allowing participants to bet on the outcome of the presidential election. Since 1988, the market has been more accurate at predicting the winner of the U.S. presidential election more often than many of the national polls have been.

In his illuminating book “The Wisdom of Crowds,” which synthesizes decades of research in behavioral economics, *New Yorker* financial columnist James Surowiecki points to other examples where a diverse, independent, and decentralized “crowd” of people made better and more informed decisions than so-called “experts.” Creating a common platform where donors and others in the sector could “bet” on some narrowly defined future state (for example, “Will the Nature Conservancy still be the largest conservation organization in 2010?”) might contribute unique analysis that “experts” at the ratings agencies could not do by themselves.

Second, a rating system needs to be able to comment on the important intangibles that are so important for nonprofit effectiveness. How is the brand perceived? How strong is the management team? Has there been any unsettling turnover of key

staff members? Is the relationship between headquarters and the affiliates healthy? Who is on the board, and how effectively do they govern? If the board is overweight with luminaries, is there a functioning executive committee that provides adequate fiduciary oversight? How transparent is the organization in reporting its finances or responding to these kinds of questions?

These are not questions that can or should be “quantified” into some simple metric, but a thoughtful analyst could fairly quickly ask and report on a nonprofit’s governance and management capacity. Frankly, these are many of the same questions that nonprofits that buy directors and officers or indemnity insurance have to answer annually to renew their policies. A ratings agency might be able to develop a revenue- or asset-adjusted measure of how much a nonprofit pays for D&O insurance, offering some more quantitative insight into how the private insurance markets value a nonprofit’s organizational integrity.

Third, ratings organizations need to address the question of social impact. We do not underestimate the difficulty of this task. Academics and leading practitioners have struggled for years with how to quantify social impact. (And, to be fair, each of the rat-



TALK BACK: Do you have a reaction to this article? Join our online forum at www.ssireview.com/forum.

ILLUSTRATION BY PHIL AND JIM BLASS/IMAGES.COM

ings agencies said that quantifying performance was one of their long-term priorities.) Unfortunately, as many organizations wrestle with how to calculate their social return on investment (SROI), some in the field are starting to question whether the methodology is too costly and complex to be meaningfully used to evaluate nonprofit effectiveness.⁸

Instead, parallels from other rating organizations might hold more short-term promise. Consider a hybrid of *Consumer Reports* or “Zagat Survey” that would conduct specific tests of the donor or service-recipient experience, and aggregate synthesized feedback from many of the nonprofit stakeholders into a reasonable (albeit qualitative) assessment of programmatic effectiveness. An elite approach would ask foundation program officers to explain why they funded certain organizations, or survey them on the perceived effectiveness of organizations that fall within their portfolios. Foundations in the United States have spent significant time and money on their performance measurement systems, and are probably as close a parallel in the nonprofit sector to the kind of for-profit financial analysts that work for investment banks.

A more democratic approach would survey donors and service recipients, or facilitate their input in a hosted environment, like Amazon book reviews or Epinions commentary. These opinions could be aggregated into a pithy user-generated synthesis, or left open source with other users asked to rank others’ feedback.

In fact, a for-profit startup in Seattle, Judy’s Book, is attempting to do just this by rating local schools and daycare centers. The company, which aims to combine local search and social networking for a range of commercial services, has launched a pilot project in Seattle to complement existing quantitative school data which, like nonprofit financial data, is also of debatable accuracy with parent surveys and feedback about schools in areas such as teaching staff, student diversity, quality of extracurricular activities, and facility conditions.

“The first thing most parents do when they start evaluating schools is to ask their friends and other parents in the community what they recommend,” said Andy Sack, CEO of Judy’s Book. “We thought the survey would make the process of getting at this type of word-of-mouth information much easier. Rather than making dozens of phone calls to friends to get a handful of responses, parents can quickly get the ‘inside scoop’ on local schools from hundreds of other families simply by reviewing the survey results on our Web site.”

While there are obvious logistical and legal questions to implementing a system like this nationally, and it is far easier in fields like higher education or medical care than homelessness or habitat protection, a ratings agency could aggregate service recipients’ perspectives to inform donor choice.

The ratings agencies could also partner with relevant domain experts to define standards of program effectiveness. High/Scope, for example has spent years studying early child-

hood education, and has accumulated great insight into the characteristics of effective childcare or preschool programs and the organizations that run them. A ratings agency could work with High/Scope to develop standards for quality childcare organizations. Admittedly, these would be qualitative standards, not quantitative outcomes, but could at least be used to indicate whether an organization is incorporating known best practices into its program design.

Need for a Business Model

One factor standing in the way of such a ratings system is the lack of a clear business model. Bob Ottenhoff noted that GuideStar has been exploring various ways to improve their analyst reports, “But for the life of me I can’t figure out the economics. The time and cost of doing this kind of research is considerable and I really doubt that donors will pay for the research.”

Perhaps a consortium of philanthropists or foundations, like the group that collaborated to start GuideStar in the 1990s will recognize the potential benefit of a new approach to ratings and invest in the needed infrastructure. Maybe existing players like *Consumers Reports* or “Zagat” (as the Better Business Bureau did), or startups like Judy’s Book, will create a market where others don’t and will start extending their own service ratings into the nonprofit sector.

Or maybe an enterprising entrepreneur will crack the code on the business model and can develop an independent and financially self-sufficient ratings system. Whatever the path, the existing nonprofit rating services deserve kudos for their initiative, but still have a long way to go before they, or anyone else, can provide meaningful guidance to donors looking to allocate scarce philanthropic dollars among various worthy causes. □

1 The three sites can be found at: www.charitywatch.org; www.charitynavigator.org; www.give.org. The Wise Giving Alliance is a project of the Better Business Bureau.

2 Wise Giving was formed in 2001 as the merger of the National Charities Information Bureau and the Council of Better Business Bureaus Foundation and its Philanthropic Advisory Service, each of which had been in existence for several years prior.

3 The one exception was the US Fund for UNICEF, which earned a C-minus from AIP, despite passing Wise Giving’s test and getting four stars from Charity Navigator. Why? When we called AIP, we were told that they exclude in-kind giving from their fundraising efficiency calculation. Since UNICEF receives millions of in-kind contributions that are primarily program-related, their fundraising efficiency ratio (fundraising expenses divided by related contributions) is downgraded.

4 Charity Navigator is a private foundation, not an operating nonprofit, so its 990 can’t be rated in the same way that it rates other nonprofits.

5 As reported in United States General Accounting Office April 2002 Report: Tax-Exempt Organizations: Improvements Possible in Public, IRS, and State Oversight of Charities.

6 The report is also available as a PDF on the Internet for online members who donate \$35 or more via Network for Good online donations service.

7 A discussion of the value of nonprofit brands can be found in Maisie O’Flanagan and Lynn Taliento’s article entitled “Nonprofits: Ensuring That Bigger is Better,” *McKinsey Quarterly* 2004, No. 2.

8 Mark Kramer, “Measuring Innovation: Evaluation in the Field of Social Entrepreneurship,” Foundation Strategy Group, April 2005. On page 22, an interview with Jed Emerson highlights some of the challenges that organizations have had in calculating SROI.