THE CHALLENGES OF CREATING DATABASES TO SUPPORT RIGOROUS RESEARCH IN SOCIAL ENTREPRENEURSHIP

PAUL N. BLOOM
CENTER FOR THE ADVANCEMENT OF SOCIAL ENTREPRENEURSHIP
FUQUA SCHOOL OF BUSINESS
DUKE UNIVERSITY
DURHAM, NC 27708
919-660-7914
paul.bloom@duke.edu

CATHERINE H. CLARK
CENTER FOR THE ADVANCEMENT OF SOCIAL ENTREPRENEURSHIP
FUQUA SCHOOL OF BUSINESS
DUKE UNIVERSITY
DURHAM, NC 27708

ABSTRACT

The purpose of this paper is to explore the challenges associated with trying to create databases that serious researchers would want to use to examine issues related to social entrepreneurship. Questions are raised about defining the units to be studied, determining what to measure, deciding where to obtain data, avoiding selection bias, obtaining responses, protecting anonymity and confidentiality, managing the database, ensuring accuracy and honesty, and creating a sustainable business model.

INTRODUCTION

The shortage of rigorous empirical research on social entrepreneurship has been lamented frequently. The case studies, story-telling, and anecdotes that have filled articles about social entrepreneurship have taken knowledge development only so far, and for greater advances to be made there needs to be data made available about the characteristics, motives, strategies, behaviors, results, and impacts of social entrepreneurs and their organizations. Data that will permit rigorous statistical analyses to uncover empirical regularities are sorely needed both to help the field gain respect and to uncover the truth about what really works to improve effectiveness in social entrepreneurship. These data can be acquired in many ways, such as through the collection of primary data by individual researchers using surveys, content analyses, observation, or examination of organizational records. But to rely on individual researchers to continue to generate their own empirical data to conduct rigorous studies is not a very satisfactory approach. Research progress is likely to be glacial if only individual data-collection is done, since this can be expensive and limited in what it can accomplish.

While individual data-collection is still very much needed and should be encouraged, the alternative approach of having multiple researchers and supportive institutions collaborate to build relevant, trusted, easily-accessible databases deserves serious attention. In business schools, fields like Finance, Accounting, Marketing, Strategic Management, and Economics have benefited greatly from having data that numerous researchers can tap into to test hypotheses and theories. Finance and Accounting have the Compustat data, Marketing has data from Nielsen and IRI, Strategic

Management at one time had the PIMS database, and Economics has loads of data from the Federal Reserve Banks and other sources.

The purpose of this paper is to explore the challenges associated with trying to create databases that serious researchers would want to use to examine issues related to social entrepreneurship. The paper addresses the following topics:

- What do we consider a social entrepreneur, or a social entrepreneurial venture, and how should we set this boundary for research purposes?
- What measures would be most desirable to include in databases? How could consensus be developed among researchers and practitioners about needed measures?
- Where could these data be obtained? Are there existing databases that contain desired measures or must new data be sought? Is it worth trying to aggregate existing databases – and dealing with the associated financial costs and political issues – or would it be more practical and efficient to generate useful databases from scratch?
- How can selection bias be avoided, so that those that are sampled are not only ones that have performed well and won awards and funding?
- For data collected for the first time, how should it be done? What methods of requesting the data and incentivizing participation are likely to be most effective?
- How should the anonymity and confidentiality of those that supply data be protected? How can databases be combined without producing anonymity/confidentiality problems?
- How will the databases be updated, refined, expanded, and validated?

- How should auditing/checking be done on the accuracy and honesty of the information provided? Who should do any auditing (e.g., accounting firms)? How would audits be paid for?
- How should data be made available to researchers? Should fees be charged? Would a service organization be needed to manage the data?
- What lessons are there from other similar efforts to create usable, living databases and what should we borrow from them?

We address these issues in the remainder of the paper, providing thoughts about how some of the issues can be resolved. However, many of the issues are very complex and will require considerable debate, discussion, and hard work to resolve. We hope this paper will help to expose the issues, engender interest, and thus accelerate the completion of this work.

SOCIAL ENTREPRENEURSHIP: SETTING BOUNDARIES

As many have noted, social entrepreneurship is a multi-disciplinary field. In the academic world, researchers interested in social entrepreneurship have brought thinking from a variety of disciplines and areas, including nonprofit management, entrepreneurship, accounting, finance, marketing, strategy, sociology, economics, public policy, and law, among many others. And some of the most prominent and widely used definitions of social entrepreneurship (Dees, 2001; Martin and Osberg, 2007), include a list of attributes (e.g., heightened accountability, or systemic impact) that can really only be evaluated once an organization or a person is subjectively under review. Simply put, our best definitions of social entrepreneurs are not objective measures.

So when considering a database for social entrepreneurship, one of the critical questions is what boundaries should be set for inclusion in the database. The two extremes are clear: at one end we could include only organizations that have been selected and vetted by selected intermediaries choosing exemplary social entrepreneurs. Using this definition alone will expose us to an extreme selection bias, as discussed below, where we only consider successful organizations. The other extreme is to include any and every organization that claims to be a social entrepreneurial venture and make the pool as large as possible. We predict the solution will lie between these options, where we work to include a larger sample of organizations aiming to achieve social change, but we select some essential criteria by which if not to include them or not, at least to define and segment them. We will not attempt to resolve this issue here, but wanted to flag this as an issue that demands thoughtful resolution, especially as we consider other data sets that may bump up against our desired intentions (such as the for-profit social ventures that are part of the B Lab or GIIRS dataset, for which each entrepreneur is already committing a great deal of cooperation time.)

WHAT TO MEASURE

As a multi-disciplinary field, social entrepreneurship has an extremely diverse research constituency, drawing on scholars with varying theoretical and methodological backgrounds. Moreover, the units of analysis these scholars tend to want to explore – i.e., the individual social entrepreneurs or the organizations they found and manage – are also very diverse, addressing a wide variety of different social problems with divergent theories of change and scaling strategies. Hence, there are literally thousands of constructs that a collection of serious researchers might want to see measured to support

their research interests. Some will want to obtain measures of the personality characteristics, leadership qualities, or socio-demographic characteristics of individual social entrepreneurs, while others might want measures of how organizations are financed, structured, managed, or marketed. Still others might want measures of the health outcomes or new jobs created by the programs of organizations, while others might want measures of the financial performance of organizations, to learn about their growth or effectiveness at "scaling."

Given the diversity of interests and backgrounds of those that would be served by the availability of databases, the notion of creating one "grand" database that would be "mined" by dozens of researchers in social entrepreneurship may be far-fetched. It might be wiser to think in terms of creating separate databases that focus on (1) individual social entrepreneurs (and only secondarily on their organizations), (2) nonprofit social entrepreneurial organizations in certain sectors (i.e., health, education, poverty-alleviation, environment), or (3) for-profit social entrepreneurial organizations in certain sectors. There could be common measures tapped in all these databases to allow some type of cross-group comparisons, but most of the analyses would probably be conducted within a single database. The downside of creating multiple databases of this character is that economies of scale of usage may not be achieved, as only a few researchers might tap each database. This might require charging too high a price to each researcher for access to the data (to cover data-collection costs).

It therefore might be preferable to start with a single database that holds promise for attracting attention from the most researchers and then seeing how the economics work out. The most likely focus for the starting database would probably be nonprofit

social entrepreneurial organizations, since there are already funders of these organizations that are trying to create databases of their grantees (e.g., Echoing Green, Schwab Foundation, Ashoka). The goal would be to reach consensus on the organizational-level variables that would be included in such a "sub-grand" database. One can imagine wanting to tap a relatively straightforward set of measures and using existing definitions from standard setting organizations when appropriate, like IRIS, the Impact Reporting and Investing Standards. These could include, for example, the number of paid full-time and part-time employees, number of volunteers, number of people served, size of overall budget, sources of revenue for the overall budget (e.g., percentages brought in by fund-raising, fees for service, government grants, ancillary business income), and division of expenses in the overall budget (e.g., percentages spent on fund-raising, service-provision, marketing, management salaries). Of course, expertise in nonprofit accounting would be needed to develop clear category or account definitions that could facilitate the entering of data into the "correct" places.

What will become more difficult is to decide what measures of social impact should be sought, since extensive variety will exist in the desired impacts. One approach that we have seen used in the past is to obtain self-assessments from organizational managers to generic, scale-type questions like: *Compared to other organizations working to resolve similar social problems as your organization, how satisfied are you with how much you have alleviated the problem? (Very Satisfied to Very Dissatisfied)* Or: *How frustrated are you with the progress you are making on the problem? (Very Frustrated to Very Encouraged).* Another approach would be to develop a customized set of impact questions more like the survey questions that B Lab asks of potential B Corporations.

There have been some discussions, especially by some groups in Europe, about trying to build consensus around a common set of impact metrics used by social impact assessment consultants and professionals. The task of deciding which would be most appropriate for nonprofit and for-profit social entrepreneurs would be substantial but highly valuable. Beyond organizational data and impact data, it may also be necessary to obtain self-reports on the strategies and tactics being deployed by the organizations, which will probably not be apparent from data on budgets, staff size, or people served. So again, generic, scale-type questions may have to be developed, asking about things like the alliances they have formed (e.g., *We have accomplished more through joint action with other organizations than we could have by flying solo.)* or their approach to replication or expansion (e.g., *We have a "package" or "system" that can work effectively in multiple locations or situations.).* To the extent that it would be possible to obtain multiple, converging self-reports on these measures from within the same organization, the data would be more reliable and useful.

Regardless of the focus of a new database (or databases), considerable work will need to be done to develop consensus among researchers and practitioners about what should be measured (as well as about how to do the measurement, as discussed below). A steering group or advisory board of talented researchers and practitioners who are likely to use the database would have to be convened and, through a process of debate and negotiation, the features of the data could be determined. This group needs to be large enough to make sure that all the important viewpoints are considered, but it cannot be too large to make convening and consensus building frustrating and cumbersome. Apparently, the Panel Study of Entrepreneurial Dynamics database at the University of

Michigan, which has been used to study commercial entrepreneurs, was developed using such a steering group (See http://www.psed.isr.umich.edu/psed/home) .

WHERE TO OBTAIN DATA

Data about social entrepreneurs and their organizations already exist in numerous places. There are foundations, fellowship and awards programs, and impact investors that have data about their applicants and grantees. There are groups (e.g., magazines, universities, corporations) that run social venture competitions or do organizational rankings that have data on their entries and candidates. There are think tanks that have assembled data from publicly-available sources like 990 tax forms for nonprofits (e.g., Urban Institute -- see information on their National Center for Charitable Statistics at http://nccs.urban.org/) and there are consulting firms that have assembled data from nonprofits on topics like their capacity-building strengths and weaknesses (e.g., the TCC Group – see information on their Core Capacity Assessment Tool at http://www.tccccat.com/). There are also operations like B Lab, which has assembled data on the business practices and social performance of for-profit social-purpose companies and, through its subsidiary, the Global Impact Investing Rating System (GIIRS), the impact of investment funds (See . http://b-lab.force.com/GIIRS/BCorpRegistration).

The problem with all these datasets is that, in most cases, the compilations were done to serve very specific data needs of certain organizations and not to serve the needs of scholarly researchers in social entrepreneurship. Hence, many of the measures that researchers would like to analyze simply are not there, and some measures that exist may

be entered in databases in ways that are hard to interpret or categorize for data analysis. For example, text-only answers to open-ended questions may have to be content analyzed and converted into either nominal-scaled or interval-scaled data in order for meaningful analysis to be done, and this could be extremely difficult and expensive.

Another potential problem with existing databases is that the measures that researchers might want to use may not be accessible because the groups that collected the data may not be willing to share the data without charging significant fees for relinquishing their "intellectual property." The groups may have also pledged confidentiality to those that supplied data in order to obtain cooperation.

Clearly, there are existing databases, like the ones being assembled by B Lab and its subsidiary GIIRS on for-profit social ventures, which should have data soon that is ready and able to be used by a significant segment of researchers – primarily because these data have been assembled in consultation with academic researchers. But databases focused on other types of organizations may not be as "research-ready," and the likelihood is strong that many researchers would prefer that the resources that would need to be spent combining databases and overcoming access hurdles be allocated instead to developing new databases that cater to their research needs more effectively. So in addition to supplying those parts of their databases that have data that are of interest to researchers and not difficult to supply, groups that have existing databases might provide greater assistance by simply using their contacts and credibility to help persuade respondent organizations to cooperate in the creation of new research-oriented databases. Guidance from a steering group could provide direction here.

AVOIDING SELECTION BIAS

It is important to have data on some organizations (or entrepreneurs) that have done well and others that have done poorly. Variation in performance, assuming some performance metrics can be agreed upon, is crucial to have in any database. Otherwise, it will be impossible for data analyses to determine the factors that have led to strong and weak performance. Essentially, you need to look for "natural experiments" in the data in order to start to understand causal relationships, since studying causation using randomized controlled trials with organizations (or even individual entrepreneurs) as the unit of analysis is not possible.

Identifying "weaker" organizations (or entrepreneurs) to include in databases and persuading them to cooperate are huge challenges. Stronger organizations are more visible and have attracted more funding and awards. They are more likely to be part of the pool of organizations that are already being included in existing databases. They are also more likely to have the slack time and resources needed to complete a questionnaire or supply data.

One possible set of "weaker" organizations to include in databases would be "runners-up" or "rejects" in grant or award competitions. While they may be reluctant to supply data to a group that did not select them, there may be ways to overcome this using certain types of appeals and incentives. Another way to identify and recruit "weaker" organizations is to advertise, hoping that both strong and weak organizations will respond. Ads could be run in magazines, newspapers, newsletters, and websites that are likely to be read by managers of social entrepreneurial organizations. Direct mail

advertising is another option, and email messages can be sent to people who have ended up on mailing lists or directories because they have attended certain conferences, subscribed to certain publications, or joined certain social networks. Another way to "advertise" would be to set up booths at conferences so that potential participants can be intercepted and asked to become part of the data collection effort. With all these approaches, the "sales pitch" to obtain participation will have to be developed carefully. Some ideas for this pitch are covered in the next section.

OBTAINING RESPONSES

If new data is to be acquired to build databases, there are numerous options that could be pursued. Ideally, when organizations are the unit of analysis, it would be preferable for multiple informants to provide data on each organization, so as to minimize bias. While self-reports and self-assessments can be acceptable, it is better if their validity can be checked against the reports of others.

Data can be supplied by having people complete data reports or respond to surveys, but the key challenge here is getting good response rates. It is important that those who respond can be viewed as representative of the population of interest, with non-response bias being minimized. Perhaps the best incentive for people to respond is so that they can obtain the new knowledge that their responses can help to produce. Thus it is important to stress convincingly that providing data will make respondents "pioneers" in helping to develop new knowledge, plus it will make them eligible for getting the first crack at new and potentially actionable results. Hopefully, respondents will find the prospect of benchmarking their progress against that of peers to be very

attractive. Perhaps respondents can be given password-protected access to certain views of the database, along with constantly updated analyses of that data. In fact, there is a consulting firm that is currently experimenting with providing its clients with this kind of constantly updated analyses of its database on nonprofit organizations (i.e., TCC Group). A model for this might be the Kauffman Firm Survey, (see http://www.kauffman.org/kfs/). They have a downloadable data file that is updated annually and they then provide restricted and customized access to more detailed data through the University of Chicago's NORC Data Enclave. GIIRS is also working through GIIRS Analytics to provide paying investor customers with access to comparative (but privacy-compliant) data that can be used to benchmark one organization against others as well as help investors seek out investments aligned with their impact goals.

Still, there will probably be a need to provide more than faster access to new knowledge or comparative data to incentivize many people – especially those from "weaker" organizations with more limited human resources – to take the time and effort to contribute information to a database. If financial support could be obtained, it might be helpful to give small grants to organizations that cooperate, at least during the first round of data collection. How much it would take to get cooperation is hard to predict, but it is the kind of guidance that might be acquired in a focus group or through face-to-face interviews. If $100 was enough and you could get 250 organizations to supply data with such an offer, it might be a good investment of $25,000. But if that kind of money was not available, then perhaps cooperation could be obtained by offering participants a chance to win a lottery for a $1,000 prize (or more). Or a promise of some other type of

gift might be effective, such as free registration for a conference or training program, free use of a new app or software program, or a new (potentially donated) electronic device. Again, research with potential participants might provide guidance on incentives.

No matter what incentives are tried, it is important that persistence be employed in trying to obtain responses. Avoiding non-response bias is important – you want to be able to say that the people that supplied data are representative of the population of interest on all dimensions, and not just people that had the time to complete a questionnaire – and so it is worth it to try several times to obtain cooperation from both those with time on their hands and those that were incredibly busy.

PROTECTING ANONYMITY AND CONFIDENTIALITY

Offering to keep information anonymous and confidential can also help to improve response rates, as respondents should be more trustful of how the information might be used. But if you are going to make such an offer, you have to live up to it, and that is no easy matter. Code numbers must be assigned to each organization and/or individual supplying data, and the codes must be stored in an electronic or hard-copy file that can only be accessed by the researchers creating the database, who have pledged to an Institutional Review Board (at the University or research organization where they work) not to reveal the meaning of the codes to anyone – with the penalty for violating this pledge being an inability to get future research approved by the IRB (and consequently an inability to get published).

Beyond the use of limited-access code numbers, it may be necessary to prohibit the distribution to researchers of certain types of combinations of data that might make it

easy for someone to figure out the identity of an organization or individual. For example, data could be supplied on all the health care organizations in the database or on all the organizations that are headquartered in the Washington, DC area. But it might not be possible to supply the data in a way that would allow one to identify the limited number of health care organizations from Washington, DC in the database. Revealing a smaller set like that could make it very easy to figure out which organization is which once other data on the organizations like number of employees or size of budgets are revealed. The best way to avoid this problem is to obtain data from a very large set of organizations and/or individuals, including large numbers of respondents from every sector and/or location. As discussed above, this may be easier said than done. Another way to avoid this problem is to have a non-transparent coding scheme for variables like sector served or location. Health might be labeled a 2 and Washington, DC might be labeled a 14, and a researcher given access to the database might only see the numbers and not know what they mean.

MANAGING THE DATABASE OVER TIME

While it would be nice if the databases had a "crowd-sourcing" quality and could be updated regularly with volunteer labor like a Wikipedia, that approach will not be feasible. Some organizational home or "gatekeeper" will be needed, not only to protect anonymity and confidentiality, but also to manage who is allowed access to the data. Rules and procedures will have to be developed about how to enter passwords, upload and download data, and report results. Decisions will have to be made about whether data reported in articles in refereed academic journals should be made available to other

researchers – as some journals require – and, if so, how the anonymity and confidentiality of data providers can be protected when "third-party" users can obtain the data.

Other decisions will need to be made about the time intervals reported and how revisions of questions and categories should be handled. One can imagine that the first dataset for a given year will start to be used by researchers and then someone will discover a glaring error in how a question has been asked or an omitted piece of information that could have been collected. Or respondent organizations could change over time, making it possible to start to answer certain questions in later time periods that could not be answered earlier. So procedures will have to be set up for making changes to the variables over time and for informing researchers about the precise nature of those changes, so that they can account for the changes in their analyses. Naturally, all this managerial activity will have to be paid for by someone (see the discussion of sustainability below).

ENSURING ACCURACY AND HONESTY

Social entrepreneurs are big thinkers and are used to putting the best face possible on their experiences and plans, which helps them attract funding. However, in submitting data to a database it is crucial that the data be as honest and accurate as possible, not becoming a wish list or an overstatement of accomplishments and an understatement of expenses. When submitting data, respondents can be urged to be as accurate and honest as possible so as to provide researchers with a better ability to uncover the "truth" about patterns in the data. Being very clear about definitions and categories should help in this regard. Nevertheless, it may prove necessary to warn respondents that random checks or

audits will be done on the data submitted. (B Lab, for example, does annual random "audits" on 10% of its certified B Corporations. Once a company is selected for an audit, all of its answers on the entire survey are reviewed and a recommended score adjustment is documented for review by an outside committee. B Lab manages this process through MBA summer interns. For GIIRS, the audit/review process is even more robust, and is being managed by Deloitte, which has the requisite global reach). Penalties for uncovering inaccuracies could vary: they could include a redress period for minor issues, the payment of small monetary fines, or if egregious, even banishment from participation in or access to the database. Perhaps an accounting firm could be persuaded to do a certain amount of audits per year on a pro bono basis or other labor could be managed for fees that would be paid by grant funding.

CREATING A SUSTAINABLE BUSINESS MODEL

While the data itself could be stored in a "cloud" somewhere at a minimal cost, the staff working at the organizational home of this database, serving as the gatekeepers and updaters, will need to be compensated in some way. There is no reason to expect that having the easiest access to the data would be viewed as compensation enough for managing the database. Grant funding from a foundation could potentially serve to get this off the ground. But once started, a sustainable source of revenue would be needed. One way to obtain revenues would be to charge small fees to researchers to gain access to the data – but this is unlikely to bring in sufficient funds. Another way to obtain revenues would be to form a consortium of consulting firms, media companies, award organizations, and foundations that would pay an annual membership fee to keep the database operating. These members would have certain privileges, like receiving a

number of hours of free technical assistance to guide them in using the database or getting early access to any articles or reports created using the data – something users of the database could be required to provide before they submit papers drawing on the data to journals. Moreover, the consortium members might be willing to pay fees to keep this database operating because the data could be used to conduct rigorous evaluations and benchmarking for their clients, awardees, and grantees. Economies of scale in data collection could be achieved in this way for these organizations, eliminating the need to do numerous independent evaluations of their portfolio organizations.

LESSONS FROM OTHER EXPERIENCES

We have discussed some of the examples we have come across and their decisions and strategies to handle the concerns covered above. In addition, we have also looked at the experience of creating the Profit Impact of Marketing Strategies database, which was launched in the 1970s. This database eventually grew to over 2600 business units from more than 400 companies, with around 100 variables tracked for each unit for 6 years. It was widely used in the marketing field and it helped to launch the separate field of strategic management (Farris and Moore, 2004). Its use in academic work eventually fizzled, but it was clearly influential. Among the insights obtained from this experience that seem relevant for social entrepreneurship are the following:

- It helps to have a respected organization, known for working with both academics and practitioners, leading the effort. In its early years, the highly-regarded Marketing Science Institute, which at that time was affiliated with the Harvard Business School (though it no longer is), fulfilled this role.

- It is beneficial to use questions that vary in response format – i.e., some should be multiple choice, other scales, other percentages – as that seems to reduce respondent fatigue and improve response rates.
- It is valuable to have a disguise factor, known only to respondents, to improve response rates (and achieve confidentiality). PIMS was set up so that most financial measures were useful only as ratios to other measures that had the same disguise factor.
- It is important to develop long-standing, trustworthy partnerships with a limited number of academics, instead of allowing any academic to use the data. If poorly-done research studies are conducted with the data, then this could damage the credibility of the entire database.

But we know there are more insights to be obtained from past database-creation attempts. We are just beginning to explore the structures other academics have used in detail in order to fine-tune our option set. We also welcome feedback and suggestions on other data collection and use models that we may not yet have considered.

NEXT STEPS

Together with the Skoll Centre for Social Entrepreneurship at Oxford University, we are starting some serious explorations around these issues. We have initiated a two-step process. Step one is to convene some prominent funding and award intermediaries in the social entrepreneurship space globally, and start to hone our ideas with their guidance and feedback about what is feasible and desirable. In a sense, they are the gatekeepers for most of the "recognized" social entrepreneurs around the world and we

want their input as well as critical knowledge about the databases they are already creating and maintaining. Step two is to convene interested academics to become clearer about the essential kinds of data that is needed and what ideas and experience researchers have in creating usable datasets from global entrepreneurial organizations.

REFERENCES

Dees, J. G. 2001. *The meaning of social entrepreneurship.* Working paper, Center for the Advancement of Social Entrepreneurship, Fuqua School of Business, Duke University, www.caseatduke.org.

Farris, P.W. and Moore, M.J. (Eds.) 2004. *The profit impact of marketing strategy project: Retrospect and prospects.* Cambridge: Cambridge University Press.

Martin, R.L. and Osberg, S. 2007. Social entrepreneurship: The case for definition. *Stanford Social Innovation Review*, Spring, 29-39.