Daniel and the rhinoceros

David E.K. Hunter *

Director of Evaluation, The Edna McConnell Clark Foundation, 415 Madison Avenue, 10th Floor, New York, NY 10017, USA

Abstract

This paper examines ways in which funders often do harm in the name of good by focusing on randomized control experiments over all other evaluation methods when helping not-for-profit organizations improve the effectiveness of their programs. It offers a critique of current practice and suggests ways in which foundations might work usefully and productively with grantees on evaluation-related capacity-building. Using a biblical example of an early evaluation, it notes how even simple evaluations that fall short of meeting the criteria of the randomized experiment can be really meaningful, useful and cost-effective for both grantee organizations and funders.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Capacity-building; Randomized experiment; Evaluation methods

1. An early evaluation and what we can learn from it

Sometimes, when confronting contemporary challenges, it is useful to consider the lessons of history. Our cultural heritage includes the story of the focused, high-stakes evaluation of a dietary program undertaken some 2600 years ago in 605 BCE. That was the year in which Nebuchadnezzar, King of Babylon, conquered Jerusalem and carried off into servitude thousands of the people of Israel. It seems that Nebuchadnezzar was a progressive ruler who wanted to assimilate his captives and turn them into productive citizens (but not too productive—he had the men castrated). He therefore commanded Ashpenaz, the master of his eunuchs, to select a group of Israelite aristocrats who were free from physical blemish, skillful, wise, and educated in science, and therefore able to comport themselves appropriately and in an aesthetically pleasing manner in his palace—so that they might be taught the language, culture, and knowledge of their hosts and then serve the king as mid-level bureaucrats. To ensure the success of this undertaking, Nebuchadnezzar ordered that they receive a daily provision of the king's own meat and wine, intending thus to nourish them for 3 years so that, at the end of this period, they might appear before him healthy, well-educated, and ready to earn their keep.

Among the selected group was Daniel as well as three others—Hananiah, Mishael, and Azariah. Whereas other

Israelites agreed to Nebuchadnezzar's terms, Daniel and his friends could not bring themselves to eat what Jewish dietary laws considered to be unclean food. He therefore demurred, threatening a hunger strike, and thereby put Ashpenaz, the newly appointed Director of the Babylonian Government Acculturation and Dietary Program for Israelites, in a terrible bind. As Ashpenaz said to Daniel, "The buck stops with me and the king is one unforgiving boss…". He went on to explain that if Daniel and his companions were to go on a hunger strike and stand before Nebuchadnezzar looking sick or emaciated, "The king'll have my head on a platter".

But Daniel, being, after all, a prophet and hence presumably able to see some 2600 years into the future, contemplated the 'Government Performance and Results Act of 1993' (about which there will be more to say later) and proposed a dietary program evaluation, suggesting that he and his friends be given 10 days to eat a porridge of legumes (thus avoiding unclean meat) and drink water according to Jewish law. And he made it a high-stakes evaluation, saying in good King James argot, "Then let our countenances be looked upon before thee, and the countenance of the children that eat of the portion of the king's meat: and as thou seest, deal with thy servants". Upon hearing this plan, Ashpenaz, through his Assistant Director Melzar, rather courageously agreed.

What happened? Well, after 10 days Daniel and his three friends looked healthier and chubbier than their compatriots who ate the king's dishes. Ashpenaz and Melzar, accepting these outcome data, allowed Daniel and his friends to follow Jewish dietary law and gave them a steady diet of pulse,

---

* Tel.: +1 212 551 9122; fax: +1 212 421 9325.

E-mail address: dhunter@emcf.org

a leguminous porridge. And further, we can assume that Ashpenaz and Melzar did not leave matters at that, and instituted a performance-tracking system through which the ongoing health of Daniel and his friends was monitored. In any event, as we all know, they thrived, and Daniel, at least, was a star performer—a standout when compared with the other Israelites who continued to eat the king's meat.

Let's take a look at this evaluation. Even though we are dealing here with Daniel, who gets the benefit of the doubt by virtue of being blessed by God and a biblical prophet, the methodology from a research perspective seems pretty weak. First of all, not only are four subjects a ridiculously small number to use for the purpose of comparison, the four that participated in this evaluation were self-selected. Furthermore, the intervention dosage was minimal; the time trial was ludicrously short; and the indicators were ambiguous. When we also consider that the outcomes were, at best, impressionistic and the measures purely subjective, it becomes clear that this evaluation was not what we would call scientifically sound. So, any evaluation novice will recognize that this evaluation lacks credibility on virtually all fronts: it has statistically useless data because of the tiny number of participants, is rife with built-in selection bias, features a weak and ridiculously short intervention, and is woeful in its lack of precision in measurement. In short, it is a mess.

But on closer inspection, we have to acknowledge that this evaluation has some virtues, too. It assessed a clearly articulated theory of change (kosher diet leads to good health—better than the king's diet), it was participatory in that program participants had a full voice in the evaluation design (actually, they proposed it), and it was led by a program director who accepted the research methodology as dispositive even though the outcomes were qualitative in nature. When we also consider that, subsequent to the evaluation and based on its findings, the program (dietary course of action) that it assessed was implemented, that an ongoing evaluation capacity was built within the program so that it could be used for continuous performance monitoring and quality improvement, and finally that it allowed for some spectacular long-term performance outcomes (Daniel, after all, correctly prophesized the fall of Babylon to the Persians), we must acknowledge that this evaluation had some noteworthy qualities.

So, here we have an evaluation that, at first glance, is problematic at best, and useless, at worst. Nevertheless, it was designed to be useful, was inexpensive, built local evaluation capacity, supported high-stakes decision-making, was used to develop an ongoing program performance monitoring and quality management system, accomplished what it was designed to do, and consequently, in fact, was useful to all involved. That is not bad at all—and worth keeping in mind as we consider the ways in which evaluation can be meaningful and useful both to not-for-profit grantee organizations and to funders—criteria that present funder-driven evaluation practice mostly honors with their absence.

2. The 'gold standard' of program evaluation: a rhinoceros in the living room

Increasingly, foundations are accepting the importance of evaluating their grants, programs, and initiatives. They do so in order to accomplish various purposes, including: (a) learning about their undertakings and what it takes to get them done, (b) assessing what they have accomplished, (c) holding their grantees and themselves accountable, and (d) developing and disseminating knowledge to their fields of interest. Yet, in spite of these perfectly good intentions, it is not unusual for significant tensions to mar relationships not only between foundation evaluators and program officers, but, even more problematically, between funders and the grantee organizations whose programs are being evaluated.

It is this latter situation that I want to think about here— because we need to get past it in order to strengthen the effectiveness and credibility of the non-profit sector and the funders who support its organizations. The stakes are high and getting higher as the federal government continues to retrench and devolve to the not-for-profit (or so-called 'social') sector its basic safety net functions. Yet, sadly, more often than not the non-profit sector is being hurt more than helped by funders' approaches to evaluation. This happens for at least two reasons. First, some funders—with more than a bit of self-interested complicity on the part of professional evaluators—have been pushed toward a very narrow focus equating evaluation with accountability and quantified outcomes, and they are imposing this view wholesale on their grantees. Second, other funders, rejecting the procrustean reductionism of the former group, are colluding with grantees, either in rejecting the value of evaluation totally or settling for such 'soft' evaluations and/or outcome measures, that their meaning is obscure at best. Both groups are, in my view, missing the most fundamental point of all, namely, that to be really meaningful, useful and cost-effective for both grantee organizations and funders, evaluation should not be thought about, nor implemented, outside or even alongside an organization's operations. Rather, evaluation should be integrated fully into organizations' operational capacities—just as Ashpenaz and Melzar did.

Consider two important developments that—though well-intentioned—helped get us into the mess we are in at the present time.

The first was the passage of the 'Government Performance and Results Act' (Government Accountability Office (GAO), 1993). This was an act "(t)o provide for the establishment of strategic planning and performance measurement in the Federal Government, and for other purposes". GPRA, as the act is known colloquially, was intended to address several very real problems, including waste and inefficiency in Federal programs, and inadequate information about their performance. The purposes of the act were, among other things, to improve the confidence of the American people in the capability of the federal government, in federal program effectiveness, and in public accountability. By promoting a new focus on results, service quality, and customer satisfaction,

GPRA was intended to improve management of the federal government and congressional decision-making by providing more objective information on whether and to what degree statutory objectives are achieved.

It is inescapable that GPRA is focused on the issue of public accountability. It is grounded in an appreciation of government's limited resources and the conviction, consequently, that evaluation must be used for improving, measurably, the results or outcomes achieved with funds spent. In other words, GPRA is concerned with efficiency as much as it is with effectiveness—hardly surprising when we look at it from a public policy perspective. It is further clear that GPRA gives primary value to quantitative data and the evaluation methods necessary to develop and analyze them satisfactorily. It requires that evaluative plans establish performance goals, express such goals in an objective, quantifiable, and measurable form (unless another form is authorized), establish performance indicators to be used in measuring or assessing the relevant outputs, service levels, and outcomes of each program activity; and describe the means to be used to verify and validate measured values.

GPRA legislation is silent on the use evaluation for purposes other than managing accountability. Yet program stakeholders most often use evaluations in other ways: (a) to learn about challenges in the design and implementation of programs; (b) monitor and learn from implementation; (c) identify, separate, and learn about the contributions of individual program elements to the achievement of outcome objectives; (d) use internal evaluation systems in an ongoing manner to manage program quality and sustain efficacy; (e) and/or capture essential knowledge about programs to inform fields of work and future program designs. This set of potential uses of evaluation systems is precisely what front-line organizations most desperately need to incorporate into their operations in order to support their ability to manage appropriately and efficiently, improve their performance (achieve targeted outcomes and impacts), grow productively, learn from their work and the work of others in their fields, and develop an empirical basis for persuading funders that they are worthy of support.

Recently, the GPRA focus on accountability has taken what is arguably an even more unreasonable turn from the front-line non-profit organizations' point of view: the federal Office of Management and Budget (OMB, 2004) has required the randomized experiment as the preferred form of evidence concerning programs. Well, is not that okay? If we are going to evaluate programs, why not rely exclusively on the 'gold standard' to design and implement our evaluations? To be blunt, no it is not. Certainly not all the time. The reason is that the exclusive focus on accountability and on the randomized experiment actually detracts from the capacity of most non-profit organizations to use evaluation meaningfully in their daily work—and can overburden their organizational capacities to the point of threatening their sustainability.

First, a quick detour back to the Bible: Daniel and his friends designed an alternative to the Babylonian Government Acculturation and Dietary Program for Israelites. Then they designed an evaluation to look at two approaches to maintaining the health and alertness of castrated captives: the King's Dietary Program and the Kosher Dietary Program. Nebuchadnezzar believed that the King's Dietary Program was essential to produce better health outcomes for participants than any other dietary regimen. Daniel and his friends believed their Kosher Dietary Program could do as well or better. In the ensuing evaluation, it turned out that the theory of causation underlying the King's Dietary Program did not hold as an alternative to the Kosher Dietary Program and could be rejected with some confidence by using a standard of 'reasonable doubt' accepted by all local constituencies.

In today's language, what Daniel and his friends argued was that if we want to understand a program's effectiveness, our methods must identify outcomes achieved by participants and ways of explaining why they happened. One explanation would be that the program is producing the expected results. But to be confident of this, we must rule out other explanations that rival the view that it is the program that is producing the outcomes we observe. These rival explanations are familiar to evaluators, and include such things as self-selection among participants, normal developmental processes that occur regardless of program participation, and statistical artifacts such as regression to the mean (Shadish, Cook, & Campbell, 2001). Short of divine revelation, of course, there is no absolute way of knowing which explanation is 'right'. When we make causal attributions, we have to settle for the every-day workable standard of 'reasonable doubt'. Sometimes it is easy to show program interventions as causative using this standard, and other times it is very difficult.

The randomized experiment achieves this standard through research design and statistical analysis. However, the randomized experiment by itself provides very little information about program elements, processes, and implementation efforts—key issues to people who run and work in programs. However, well-designed randomized experiments usually are expensive and time-consuming. Other problems of randomized experiments have been amply discussed in the research literature for decades, such as the need for replication and the problem that a statistically significant result may be trivial at a practical level. Rather less discussed, is the tendency for accountability studies to be conducted by outside experts, adding necessary objectivity but providing little of value to practitioners' self-assessment and quality improvement capacities. With these points in mind, the question gets changed from "Why not use randomized experiments in all program evaluations?" to "When does it make sense to use randomized experiments and when are other designs preferable?"

Although OMB recognized that randomized experiments could not be conducted in every program, overall the focus on this method has proven a procrustean bed. The GPRA approach does not appreciate the extent to which program context demands other designs, and it fails to meet the needs of those seeking answers to evaluative questions beyond the matter of impact. So there is reason for funders to think twice before requiring such evaluations of their grantees.

On the other hand, as was mentioned above, there are funders and non-profits that far too uncritically buy in to these

criticisms and categorically reject randomized experiments and accountability-based approaches to evaluation so fully that they throw out the proverbial baby with the bathwater. They see evaluation costs as threats to direct service resources, disparage the 'science' of evaluation as inappropriate to the 'art' of grantmaking and the work of grantees, and at best will sign on to process or formative evaluations that can not in any way be used for purposes of assessing effectiveness or accountability. Adherents of these views nonetheless often advocate strenuously for favored programs, initiatives, and approaches to changing the world for the better, and often do so effectively—thereby channeling scarce resources into support for untested activities that may or may not accomplish what they set out to achieve (and with significant opportunity cost)—or even worse, may do harm. [Evaluators will know many examples of programs that do more harm than good; one is Scared Straight, which exposes youngsters headed toward criminality to violent felons in prisons in order to scare them back toward the 'straight and narrow'—and instead apparently motivates them to make themselves tougher and more inured to compassion than they were before, while doing nothing to move them off a trajectory toward criminality and incarceration (Finckenauer, 1982)].

So, in summary, we should not be surprised that, when the use of evaluation by foundations is assessed, the picture is dismal. Frequently, even where foundations commission evaluations, the reports are not used by program staff to inform grantmaking decisions. When asked why this is so, foundation staff have complained that outcome and impact evaluations are too late to be of use, too academic to be helpful, too crude to be enlightening ('proving the obvious'), too dense and methodologically burdened to be interesting, and too expensive to boot. Grantees, in turn, often find funders' evaluation requirements exceedingly burdensome, of no practical value, and, in fact, a drain of precious resources that threatens their ability to do their work (Patrizi & McMullen, 1999).

This need not be so. The value that evaluation can bring to grantmaking—that is, to the work of funders and grantees alike—is clear, and clearly demonstrable. But this will only hold if we learn from the Babylonian Government Acculturation and Dietary Program for Israelites evaluation that established the Kosher Dietary Program's effectiveness and thereby substantiated its value to Nebuchadnezzar the funder, Ashpenaz the program director, and of course to Daniel and his friends, the program participants. In doing so we need to drive the GPRA/OMB rhino out of our living rooms (the realm of the quotidian) and into the confined preserve (of special cases) where it belongs.

3. Corralling the rhinoceros: building evaluation capacity at the grantee level

The Nebuchadnezzar evaluation, let us recall, embodied some notable virtues. It was designed to be useful, was inexpensive, built local evaluation capacity, supported high-stakes decision-making, was used to develop an ongoing program performance monitoring and quality management system, accomplished what it was designed to do, and, consequently, in the end was useful to all involved. It is possible for modern program evaluation efforts to do as well, and the Edna McConnell Clark Foundation's approach to evaluation-based work with its grantees can illustrate this.

The Edna McConnell Clark Foundation searches for, selects, invests in, and supports high-performing youth-serving organizations (with a priority on those who serve older youth from low-income backgrounds) that have the potential to grow in capacity or scale—through large, multi-year capacity building grants and associated non-financial support. The Foundation has learned that grantees benefit from consultations provided in the area of evaluation, in which they are assisted in specifying the group(s) they seek to serve, clarifying outcome objectives for programs participants, describing program elements through which they intend to help participants achieve targeted outcomes, and identifying the human, material, organizational, and fiscal resources needed to deliver services as intended.

This amounts to developing a theory of change—a formal rendering of the approach adopted by the organization to change something about the world. The theory of change becomes the guide whereby the organization structures its daily activities to achieve its strategic goals and objectives. It also provides the framework within which each organization can examine what works and what does not work within its own programming, and manage performance for continuous improvement. Companion papers in this special issue of Evaluation and Program Planning describe methods we have developed to address these issues as well as suggestions to help grantees develop strong theories (Hunter, 2006). Evaluation staff at the Foundation also consult with grantees regarding the construction of their in-house evaluation capacities. In doing so, we have identified the following three conceptually consecutive—but, in actuality, overlapping—developmental stages for building such capacity.

3.1. Stage 1: participant profile

This refers to the capacity to know details about whom one is serving. At the very least, an organization should know the name, residence, age, gender, ethnicity, and socioeconomic status of every program participant. For youth-serving agencies, it is also essential to document basic information about the family. And, organizations wishing to learn about what outcomes clients achieve, must of course develop baseline data—which means capturing appropriate information about how each new client rates on key outcome indicators that will be used to measure programmatic success—during program participation and also after he or she leaves the program. Implementing this most basic phase of an internal evaluation system can take several years.

3.2. Stage 2: participation patterns

Participation information provides a description of 'how much' of a program the participants actually get. These data

are essential for monitoring and improving program quality. For instance, information about program participation patterns can help organizations identify sub-groups who use their programming less than expected or who drop out altogether—and suggest ways to address such issues. Also, participation data can be combined with outcome data to look at whether the participants are getting 'enough' services to produce desired outcomes. Organizations can not really engage in ongoing program quality management—let alone improvement—without systematically collecting participation data. In fact, it is not very useful to look at outcomes until there is confidence that participants do receive the level of services intended. Implementing this evaluation capacity often takes another year or two after an organization has implemented its participant profile database.

3.3. Stage 3: participant outcomes

Documenting information about the progress and outcomes of clients is the basis for assessing whether program participants actually are benefiting. Most people think of this evaluation component as answering the question, "Does the program work?" Actually, collecting outcome data without comparative evaluation efforts (to rule out alternative explanations for the achievement of outcomes) will not answer this question. However, exploring data about participants' progress and outcomes can be very useful in helping organizations think about other questions, such as "Does the program apparently work better for some participants than for others?" or "Are some outcomes easier to achieve than others?" Here again, an organization's evaluation capacity becomes the basis for program management and improvement as well as the basis for—ultimately—undertaking external, comparative evaluations to assess how effective its programs really are. Designing and implementing the indicators, measures, and methods through which outcomes are assessed, also can take several years.

Thus, long before it is appropriate to undertake an experimental program evaluation, building an internal evaluation system is essential for an organization to manage its programs well and learn sufficiently from its operations to assure the maintenance of high quality programming. Consulting to grantees about the design, implementation, and maintenance of such systems is of great value to those that are committed to being effective—that is, serving their clients well. This work, then, quintessentially grounded in evaluation knowledge and practice, is intrinsically useful and important. Such consultations can be undertaken directly or supported via technical assistance providers, by funders large and small.

When does it make sense to undertake external comparative evaluations? When an organization has reached sufficient developmental capacity to offer its services reliably, at a high level of quality, to a significant number (100s) of service-recipients, with high levels of participation and low levels of premature drop-out, over an extended period of time (years),

with a full range of evaluation data having been collected, and within a secure and sustainable institutional setting.

However, even at this point, a randomized experiment might not be needed. Alternative evaluation methods can answer important questions about program effectiveness that help us rule out other causes for outcomes gained by participants. These include the wide variety of quasi-experimental designs, case study designs, and meta-analyses. All of these approaches require that evaluators have knowledge about the range of possible causes of program participant outcomes (such as children maturing developmentally and thus 'naturally' acquiring new skills and competencies). Clearly, the more codified and focused the program, the more discrete the outcome that one is assessing, and the greater the number and kinds of independent data sources one uses—the more likely it is that such methods will illuminate causality and hence program impact.

The point is not to claim that such approaches provide ultimate answers to program effectiveness—randomized experiments do not do that either! Rather, the idea here is that these methods provide very useful information for understanding what programs really are doing, how they work, and how effective they likely are in helping participants achieve key outcomes.

Of course, there may come a time to put the rhinoceros to work—when it may make strategic sense for a given organization to make use of what has become the widely accepted 'gold standard' of evaluation design (use of RCT methods) to confirm in ways generally regarded as scientifically rigorous that a program is working as intended for participants—that is, achieving measurable impact. Doing so can have splendid results—note, for example, the allocation of federal funding to the field of mentoring after the P/PV evaluation of the Big Brothers Big Sisters mentoring program (Grossman & Tierney, 1998). (Of course, even when this has been done at one location, it would be unwise to assume that impacts will necessarily be achieved in a like manner at other locations with other staff and other participants.)

4. Summary

Funders, often acting on the advice of evaluators, can be heavy-handed and arbitrary in imposing evaluation requirements on grantees, and this can be harmful. Evaluation need not be burdensome or expensive to be useful, especially when it comes to developing in-house evaluation capacity. There is great value in a simple, step-wise approach such as the one discussed here—and I can report with confidence (based on anonymous, yearly interviews of all the Foundation's grantees conducted by an external evaluator) that the Foundation's grantees agree. There is a terrific payoff to working with non-profit organizations to help them build their evaluation capacities as part of learning about 'what works'. If we do so, long before we can be 'sure' about 'what works' we can be confident that such organizations will be busy assessing key issues of program quality, making ongoing efforts to improve quality and effectiveness, and consequently

will be much more likely to be accomplishing every day what they have set out to do.

References

Finckenauer, J. O. (1982). *Scared straight and the panacea phenomenon*. Englewood Cliffs, NJ: Prentice-Hall.

Government performance and results act of 1993. Washington, DC: Government Accountability Office (GAO). (1993).

Grossman, J. B., & Tierney, J. P. (1998). Does mentoring work? An impact study of the big brothers big sisters program. *Evaluation Review*, *22*(3), 403–426.

Hunter, D. (2006). Using a theory of change approach to build organizational strength, capacity and sustainability with not-for-profit organizations in the human services sector. Evaluation and Program Planning doi: 10.1016/j.evalprogplan.2005.10.003

What constitutes strong evidence of a program's effectiveness? *Revised program assessment rating tool (PART) guidance*. Washington, DC: Office of Management and Budget (OMB). (2004).

Patrizi, P., & McMullen, B. (1999). Realizing the potential of program evaluation. *Foundation News and Commentary*, *40*(3), 30–35.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.