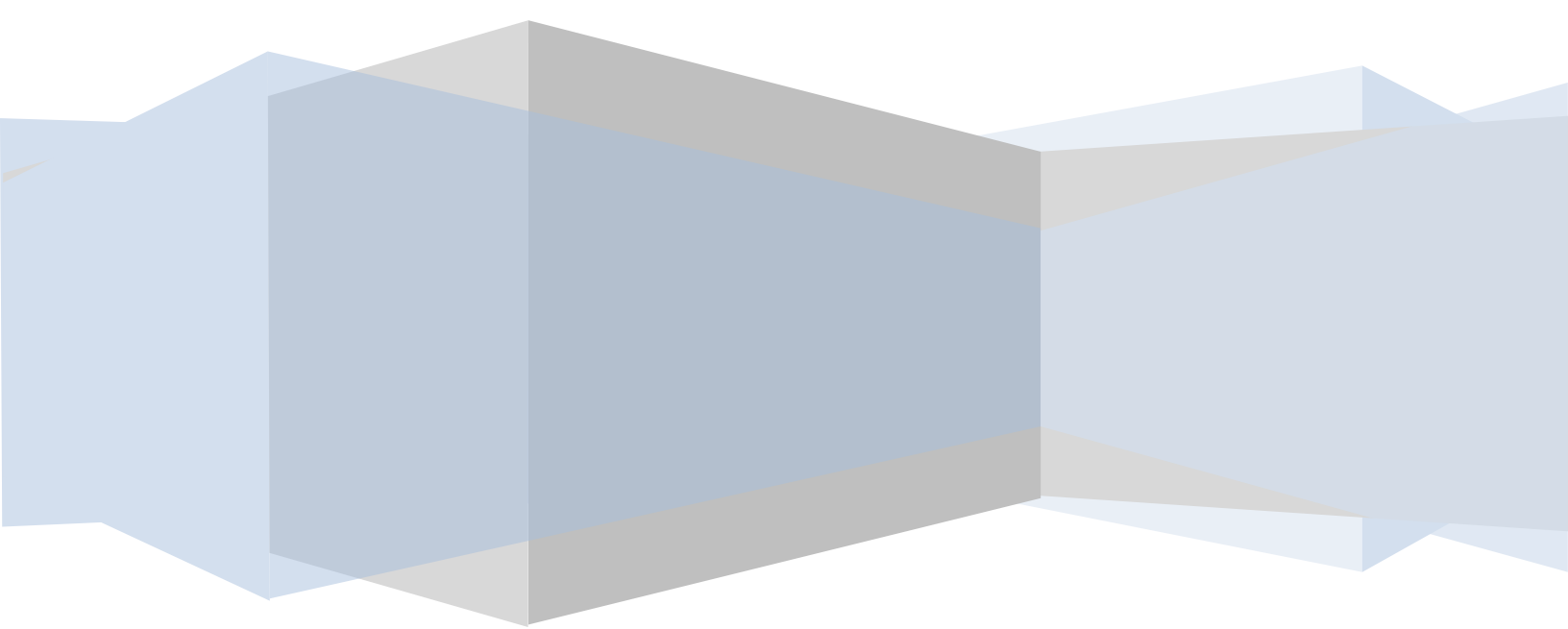


Paper presented at the ISTR Conference, Barcelona, Spain, July 9-12, 2008

# Program Evaluation for Accountability: What's Different Now?

**Richard Hoefler, Ph.D.**  
**Professor**  
**School of Social Work**  
**Box 19129**  
**University of Texas at Arlington**  
**Arlington, TX 76019 USA**  
[rhoefler@uta.edu](mailto:rhoefler@uta.edu)

**Abstract:** This paper examines the status of the use of program evaluation in human services nonprofits in the Dallas, Texas, USA area. Using methodology and questions similar to Hoefler (2000) and comparing answers over time, the results indicate that attitudes and use of program evaluation to ensure accountability is increasing. Despite improvement, agencies fall short in some areas, particularly in terms of communicating results to persons outside the agency.



## **TITLE: Program Evaluation for Accountability: What's Different Now?**

Dr. Richard Hoefler  
University of Texas at Arlington

### **Introduction**

The cry for greater accountability within the nonprofit world is a constant background noise for leaders within the sector (Bailey, 2005; Cohen, 2006; Lee, 2004, Walden, 2006). Waves of scandals have broken over the field, tearing down funding prospects for organizations and programs, good and bad alike (Carson, 2002; Stephenson, 2006). Even faith-based organizations are called upon to provide greater evidence of effectiveness through program evaluation (Fischer & Stelter, 2006).

Leaders in the nonprofit world attempt to blunt criticism by showing the efforts made to assure greater accountability in providing information on the ways they behave (activities), what they produce (outputs) and, especially, their achievements (outcomes), as they justify costs while seeking operating funds (Houchin, 2002).

In the United States, for example, nonprofits increasingly place their financial and programmatic information on websites, even as those same websites are re-tooled to become more effective fund-raising vehicles. International organizations are also moving forward with communication campaigns that seek to reassure potential donors that their contributions are well-spent, as accountability is certainly an issue in the third sector beyond the US context (Ferguson, 2005; Flanagan, 2007). Journal articles call for improvements in the specifics of ensuring accountability in, for example, Australia (Flack & Ryan, 2005) and Canada (Burnley, Mathews & McKenzie, 2005).

In some quarters, a backlash against the calls for accountability has emerged (Benjamin, 2008; Ebrahim, 2005; Irvin, 2005). These authors argue that “hyper-accountability” is decreasing nonprofits’ ability to actually achieve their desired outcomes.

Most of the literature is based on case examples or theoretical approaches to accountability. This paper takes an empirical look at what agencies are doing in terms of the use of evaluation, linking it to a particular approach to what agencies should do to use program evaluation for maximum accountability. It uses the results of a survey conducted first in 1999 and repeated in 2008, to examine if the use of evaluation is changing, if so, in what ways, and whether accountability is being enhanced by these changes.

### **Literature Review**

Program evaluation is a frequently espoused tool for assuring that organizations are achieving their aims (as almost any text on evaluation will claim: See, for example, Rossi, Lipsey & Freeman, 2004, p. 18: “...any evaluation...worthy of its name must evaluate—that is, judge—the quality of a program’s performance as it relates to some aspect of its effectiveness in producing social benefits.”). The benefits of accountability and the use of program evaluation as

a way to assure it, are promulgated both as a normative judgment and as an empirical fact. Among other positives of using program evaluation for accountability, programs are supposed to become more effective; with more unified and focused staff, administrators and board; higher morale among workers; and more satisfied clients (for example, see Chisholm, 1995; Frumkin, 2001, Martin & Kettner, 1997; Rossi, Lipsey & Freeman, 2004).

Still, there is some question as to whether evaluation is being used as a true accountability mechanism, or if its use is primarily symbolic, at best (Tassie, Murray, & Cutt, 1998). If funders do not require action based on results, then evaluation is not being used to increase accountability.

Hoefler, (2000) lays out four criteria to determine if evaluation is being used to assure accountability:

- 1) Is evaluation being done?
- 2) Are appropriately rigorous research methods being used?
- 3) Are front-line staff members learning the results of the evaluation? and
- 4) Are other stakeholders (including the public) learning the results of the evaluation?

Hoefler suggests that unless all four criteria are met, program evaluation is not living up to its potential as a tool for organizational accountability. His study of Dallas, Texas, nonprofits suggested that accountability was not being achieved at that time in that location, due primarily to a lack of rigor in the evaluations being conducted, and the lack of communicating with stakeholders what the results of the evaluation were (Hoefler, 2000).

Other authors agree that accountability on the part of nonprofits requires greater communication by nonprofits with external stakeholders. Lee (2004) states that nonprofit organizations should consider the public-at-large as a stakeholder to which they owe a degree of “translucency,” if not “transparency” (p. 177). In order to accomplish this aim, Lee suggests nonprofits use both an indirect approach (such as using the media to generate news stories) and a direct approach (using financial and annual reports and posting this and additional information on the organization’s website). Roberts (2002) proposes the use of dialogue to keep public officials accountable, a mechanism that would presumably work to keep nonprofits accountable as well.

Brown and Troutt (2004) describe a situation in Canada where a cooperative approach to accountability was used between the provincial government and nonprofits, to good effect. Brown and Troutt (2007) indicate that reporting requirements on nonprofits may be either busy-work and difficult to accomplish, thus interfering with the nonprofit’s mission, or extremely helpful, depending on how funding from the government is set up. If too much emphasis is put upon reporting, organizational and governmental resources are wasted.

Several researchers have found that accountability requirements do have an impact on the behavior of nonprofit boards and staff. Barman (2002) notes that nonprofits engage in a process of differentiation between themselves and other organizations when they engage in a competition for resources. In this process, nonprofits choose criteria of differentiation on which they outshine the competition. While not specifically mentioned by Barman, one possible source of

differentiation is to show the use of program evaluation to demonstrate stronger program effects and accountability, and to increase an organization's legitimacy and deservingness.

O'Regan and Oster (2002) show that greater government funding impacts the way that nonprofit boards use their time, specifically by increasing the energy devoted to financial monitoring and boundary-spanning activities. Disturbingly, though, O'Regan and Oster further report that boards of organizations with greater levels of government funding also self-report themselves as being more passive than boards of organizations that are less government-funded.

Another approach to understanding nonprofit accountability is to blend the concept of performance evaluation and risk. Accountability, in this view, can be better understood "by considering how...accountability systems distribute risks in a given set of relationships and the consequences of such arrangements for nonprofit practice" (Benjamin, 2008, p. 660). Of vital concern is the possibility that "performance accountability requirements may conflict" with nonprofits getting their job done (Benjamin, 2008, p. 978).

Marvel and Marvel (2007) indicate, however, that nonprofits may not need to have great concern regarding government requirements for accountability. They show that governments tend to monitor services provided by internal employees almost as much as when those services are provided by for-profits. But, they "find strong evidence that performance monitoring by the contracting government does not extend to nonprofit and other governmental service providers, each of which is monitored much less intensively than when comparable services are provided internally. For such service providers, it appears that monitoring is either outsourced along with services, or simply reduced" (p. 521).

## **Research Questions**

The key empirical questions addressed in this research are:

- What are nonprofit human service agencies in the Dallas, Texas, area doing in terms of program evaluation? Are they conducting evaluations, and, if so, what types, using which designs and data sources, and for what purposes?
- How do these evaluation practices compare with data collected in 1999?
- What are the attitudes about program evaluation currently, and compared to 1999?
- Are there differences between agency use of program evaluation or other accountability strategies, depending on the source of funding?

This empirical information will then allow us to explore these more theoretical questions:

- Are program evaluations being used to promote accountability more now than in 1999?
- What are the reasons for any changes seen in the two data collection times?
- How do nonprofit leaders choose and implement accountability practices as they attempt to balance the multiple demands on their agencies?

## **Methods**

Survey research techniques were used to gather data from human service agencies in the Dallas, Texas metropolitan area. Agencies selected for inclusion in the sampling frame were those

included in the “Blue Book” of agencies that is compiled annually by the Community Council of Dallas. This book is a listing of agencies by type of services offered. It is the primary source of referral sources that agencies in the Dallas area use. An initial list of agencies was selected based on whether the agencies provided direct services to individuals. Many agencies were excluded from the sampling frame because they were information-providing or medical-services based agencies, rather than direct human service agencies. Examples of excluded agencies in these categories include the Arthritis Foundation, Dallas Chapter, and the American Diabetes Association. Services sponsored by city or county government agencies were also excluded in order to focus on nonprofit organizations. In all, 67 agencies were included in the final list of agencies. This is a smaller number than was used in the survey from 1999, even though the selection process was similar. After two contacts mailed contacts, the response rate is 48%, with an  $n = 32$ . Quantitative data analysis was conducted using SPSS.

## **Results**

Results indicate that change is occurring. It should be emphasized that the number of usable responses in 1999 was 91 (response rate of 66%), but is 32 in the 2008 survey (response rate of 48%). Thus, caution should be exercised in making too much regarding the comparisons made below.

### *Characteristics of the Respondents as Individuals*

- The information seems to come from similarly placed sources in the two surveys. In 1999, 89% of respondents were the executive director or president of the organization; and 3% were other top administrators. In 2008, 69% were the executive director or president of their organization with 25% being another type of top administrator. The mean length of service at the organization of the respondents in 1999 was 8.4 years, with a mean length of service in their present job at 6.9 years. This compares to the 2008 respondent figures of 7.9 years in service at the organization and 7.5 years in the current position. The mean age of respondents in 1999 was 49.2 years compared to 52.0 years of age for respondents from 2008. None of these differences is statistically significant.
- In 1999, 63% of the respondents were female (37% male); in 2008, 75% were female (25% male). (This is not a statistically significant difference.)
- In 1999, 82% of the respondents were Anglo, 10% were African-American, 6% were Hispanic, 1% was Asian and 1% was “Other”. In 2008, 84% of the respondents were Anglo, 9% were African-American, 3% were Hispanic, and 3% were Asian. None were other. This distribution is markedly unchanged over the time period in question.

### *Characteristics of the Respondent Organizations*

- The median agency budget in 1999 was \$ \$1,000,000, while in 2008 it was \$1,950,000. Organizational budget percents in 1999 were 32% from government sources, 7% from United Way, 13% from foundations and 48% from others. These percentages in 2008 were: 40% from government, 8% from United Way, 17% from foundations and 35%

from other. While the data show that government and foundation funding sources are more important now than nearly a decade ago, the differences are not statistically different.

- The average age of the organizations responding to the survey in 1999 was 39.2 years, with a range between 11 and 163 years since founding. The mean age of responding organizations in 2008 was 40.9 years, with a range between 7 and 119 years. These differences are not statistically different.
- A question asked in 2008 but not in 1999 was whether the organization considered itself “faith-based.” About one in eight (16%,  $n = 5$ ) did so. This number is too small to conduct any separate analyses on.

In general, then, we can see that the respondents are well-placed and seasoned enough in their jobs to give accurate information. Further, while there are some small differences between the sets of responding organizations, the key variables of organizational age, size and source of funding are similar in the two groups of respondents, separated by 9 years.

#### *Prevalence and Aspects of Evaluation and Its Use*

Many questions were asked that show a trend in the increased use of evaluation, and greater rigor in the types of evaluation used and the types of measures employed in the evaluations. Few of these are statistically significant, on a question-by-question basis, yet an important general trend appears clear. A greater percent of agencies are using evaluation and the evaluations may be at least somewhat more scientifically valid. **(Unless otherwise stated, differences between 1999 and 2008 responses are not statistically significantly different.)**

- *A greater percentage of agencies is conducting evaluations.* Three fourths (76%,  $n = 69$ ) of respondents had had their largest individual program evaluated within the past two years in 1999, with 84% ( $n = 27$ ) having done so in 2008. A second way of measuring this concept is to ask respondents to disagree or agree with the statement “Evaluations are conducted regularly in this organization”. On a scale of 1 to 7, with 1 indicating that the respondent very much disagrees with the statement and 7 showing very much agreement, results show considerable increase in the regularity of program evaluation. In 1999, the mean response was 5.08. This increased to 5.44 ( $p < .000$ ) in 2008.
- *The reason why agencies are conducting evaluations is changing significantly. This may also be seen as evidence of greater institutionalization of the evaluation processes within agencies.*
  - In 1999, agencies used evaluations to control program operations (this is, ensure that the program complied with proper procedures) more than any other reason (57%). This was basically unchanged in 2008 (56%). The three other reasons for conducting evaluation increased significantly during this time. Curiosity, the desire to know how the program is doing, was cited by 51% in 1999 and 80% in 2008 ( $p < .05$ ). Coercion, being required to conduct an evaluation by a funder,

increased from 39% in 1999 to 80% in 2008 ( $p < .001$ ). Finally, a commercial reason, to show funders and potential funders how successful the program is, also grew from 42% in 1999 to 80% in 2008 ( $p < .001$ ).

- *Of the organizations that conducted an evaluation in the prior two years,*
  - 66% conducted implementation monitoring in 1999; 74% in 2008.
  - 66% conducted process evaluation in 1999; 85% in 2008;
  - 82% conducted an outcome evaluation in 1999; 93% in 2008.
  
- *Evaluation may also be becoming more institutionalized, both in terms of staff knowledge and funding.*
  - The sources of expertise for the evaluation in 1999 were: 13% inside the organization, 16% outside the organization and 71% from both inside and outside the organization. In 2008, the sources were 33% inside the organization, 15% outside the organization and 52% from both inside and outside the organization. It thus appears that more agencies are relying more on their own expertise to design and conduct their evaluations. The sole use of outside expertise is down markedly.
  
  - Funding for evaluation is now more likely to come from internal agency funds or be part of the grant that funds the program. It is less likely to be based on donated labor. In 1999, internal agency funds were used in half of evaluations (50%). This increased to 63% in 2008. Almost one-third (29%) of evaluation efforts were funded with money from the program grant. In 2008, this had increased to 44%. In 1999, 25% of evaluation efforts relied on donated labor, while in 2008, this had dropped to 11%. Separate grants to fund evaluations were used by 13% of respondents in 1999 and 11% in 2008. (Note: more than one type of funding could be checked, so totals add to more than 100%).
  
- *It appears that agencies are slightly changing the type of design used in their evaluation. There is an important decrease in the use of the posttest only design, a greater use of the single-group pretest-posttest design, and slightly less use of a time series design. There also appears to be a retreat from the use of a comparison group design of any type. These results are mixed in terms of whether more rigorous designs are being employed.*
  - The types of design used in the different survey responses: in 1999, posttest only was used by 26%; pretest-posttest by 49%, time series by 34%, and any type of comparison group, by 15%. In 2008, posttest only was used by 15%, pretest-posttest by 59%, time series by 33%, a comparison posttest only by 4%, a comparison group pretest-posttest by 0% and any type of experimental design by 7%. (Note, this question's wording was altered to allow for finer distinctions in the more rigorous evaluation designs.)

- *Data sources may be becoming more rigorous, with the use of standardized instruments up sharply and the use of non-standardized instruments also higher. The major source of data remains agency records, by a large margin.*
  - Data sources used in 1999 were: agency records (87%), standardized instruments (34%), non-standardized instruments (54%), and social indicators (15%). In 2008, the sources were: agency records (85%), standardized instruments (52%), non-standardized instruments (63%), and social indicators (19%).
- *More agencies intend to use the evaluation results for all of the purposes listed. This seems to indicate that the agencies conducting evaluations are doing more with the results in 2008 than in 1999.*
  - The planned uses of the evaluation results in 1999 were: to make improvements in the program (96%), to verify program outcomes (77%), to shift funds away from the program (0%), to shift funding to the program (16%), to advocate for more funding from the original funder (47%), and to advocate for more funding from a new funder (59%). The planned uses in 2008 were: to make improvements in the program (100%), to verify program outcomes (89%), to shift funds away from the program (7%), to shift funding to the program (22%), to advocate for more funding from the original funder (74%), and to advocate for more funding from a new funder (78%).

While these results are preliminary, there is hope in them that evaluations are becoming a more established and recognized activity, with real uses. They also seem to be becoming at least slightly more rigorous, with the use of somewhat better designs and data sources.

#### *Attitudes about Evaluation*

In addition to the changes in behaviors noted above, attitudes about evaluation are improving. Thirteen statements were presented to the respondents who were asked to rate how true they were on a seven point scale, with a low score (1) indicating that they very much disagreed and a high score (7) indicating that they very much agreed. (Shaded rows indicate significant changes in responses from 1999 to 2008.) Four of these statements can be seen as positively regarding evaluation (*shown in italics in Table 1*), while nine can be said to be negative (**shown in boldface in Table 1**).

All four of the positive statements received a higher mean score (indicating greater agreement) from respondents in 2008 compared to 1999, two of them to a statistically significant extent. Seven of the nine negative statements got lower scores (meaning less agreement) in 2008 compared to 1999 (three at a statistically significant level). The other two statements were within .03 points of each other. Perhaps the most striking single change was the decrease from “slightly agree” to “fairly much disagree” that “evaluations are usually biased or inaccurate”.

Combined with generally more agreement about the positives of evaluations, the decreased agreement with negative statements about evaluation indicates an improved attitude towards the utility and practice of evaluation. This is certainly welcome news.

Additional information from the 2008 survey, based on questions not asked in 1999, give us room for additional thought.

Respondents were asked who was the final arbiter of how much of the evaluation's conclusions and recommendations were released and to whom. Nearly three-fourths (74%) answered that the Executive Director made the final decision. In the remainder of the agencies (26%), the organization's Board of Directors made the final decision. Another possibility was listed, that of the decision-maker being the program evaluator. No organization leaves the decision as to what to release up to the evaluator.

Once the decision was made regarding which information evaluation to release, respondents were asked to whom the information was shared. Almost all (96%) shared the information with the agency's board of directors. A large majority (85%) said that managers or administrators other than the program manager formally received information on the results of the program evaluation. The same percentage shared the information with funders and with front-line staff.

Two other possible stakeholders were asked about. The public was informed via posting on a website or publication in another form, by 39% of the organizations. The media were alerted by only one-fifth (19%) of the organizations.

Another set of questions addressed in this research is whether funding sources impact use of evaluation or accountability strategies. This was addressed two ways, first looking at whether the receipt of government funding made a difference, and then looking at whether the receipt of United Way funding made a difference. Results show that there are not significant differences between agencies that receive more than the median percent of their budget from government sources (25%) and those that receive less than the median percent of their budget from government.

The distribution of agencies receiving United Way funding differed significantly between 1999 and 2008. In 1999, only 30% of respondents received United Way funds, compared to 51% in 2008 ( $p < .05$ ). Holding the year of the survey constant, the receipt of United Way funds was associated with a significantly increased use of evaluation ( $B = .646$ , standardized Beta of  $.183$ ,  $p < .05$ ). Still, the year of survey had a stronger effect in this regression analysis ( $B = 1.244$ , standardized Beta of  $.299$ ,  $p < .001$ ). (Receipt of United Way funds was dichotomous because United Way requirements for outcome evaluation apply either totally if funding is received or not at all.)

## **Discussion**

What do these results tell us about what has changed since the end of the 20<sup>th</sup> Century, at least when considering the use of program evaluation to achieve nonprofit accountability? Initial results indicate progress is being made in the use of evaluation by human service agencies, at

least in the Dallas, Texas area. Research plans include discussions with agency respondents to collect qualitative information regarding the decisions made with respect to evaluation in agencies and its uses.

Hoefler (2000) noted problems with the use of program evaluation for accountability nearly a decade ago in that, while evaluations were being conducted, they were not particularly rigorous. Agencies seemed to be conducting evaluations primarily to assure that programs were being conducted according to plan, but not focusing enough on assuring positive changes for clients. This was indicated by the type of data sources employed. Finally, it was difficult to determine if front-line staff and other stakeholders, such as the public, actually learned of the results of the evaluations. In this way, the public was not able to hold the agencies accountable for results. Funders, such as United Way, stepped into the breach, in some cases, and demanded that agencies conduct at least a cursory evaluation using agreed upon client outcomes.

In recent years, the push from federal government funders for outcome information (spurred by the passage of the Government Performance and Results Act) has intensified, and United Way volunteers in the Dallas area have been further trained and requiring greater efforts by agencies to show adequate evaluation efforts. In at least one case known to the author, one well-known agency in the area received no funds at all from United Way after failing to improve on its evaluation activities after being warned to improve its outcome evaluation efforts for several years.

Using Hoefler's (2000) four-fold minimum requirements for the use of program evaluation for accountability, where do we now stand, comparing data from 1999 to data from 2008?

*A) Is evaluation being done?*

More so than before, evaluation is becoming standard practice. Most agencies regularly perform evaluation on at least their largest program.

*B) Are appropriately rigorous research methods being used?*

The answer to this question is not as clearly positive. Evaluations seem to be moving towards the simple pretest-posttest design as a default. While this design has many well-know problems, Mindel and Hoefler (2006) defend it as being perhaps the best possible design to expect, given agency constraints. They recommend the use of effect size analysis and replication to overcome the problems inherent in the design, while still leading to generalizable knowledge of reasonable quality.

The increased use of standardized measures as a source of evaluation data is a hopeful sign. The nature of these measures means that program impact on clients can be determined with considerably improved reliability and validity. Programs can also be more easily compared one to the next, in terms of how much client change they are producing.

*C) Are front-line staff members learning the results of the evaluation?*

According to respondents, this is happening to a great extent. About 7 out of every 8 organizations say they are feeding evaluation results to the direct service workers in their agency. Also, and perhaps connected, the attitude about how useful evaluation is to direct service

workers has improved significantly, so that there is much less agreement that results are of no use to them.

*D) Are other stakeholders (including the public) learning the results of the evaluation?*

The final element of Hoefler's model stresses the need for other stakeholders, including the public, to learn the results of the evaluation (others agree with his basic point, such as Brown & Troutt, 2004; Lee, 2004; and Roberts, 2002). Such openness to outside scrutiny is not happening to the extent that would be desirable. While nearly 40% of organizations make their evaluation information available, at least to some degree, on their website or through other publications, only a relative handful (19%) provide such information to the media. Organizations are missing out on a wonderful opportunity to press their case in public and to the public, and the public is missing out on its rightful place as watchdog over agency efficiency.

More troubling, in some ways, is the way that organizations hold on to the decision regarding which information to release. While it is understandable that evaluation reports should not be allowed to be disseminated by just anyone, it is also not in the interest of the public if evaluation results are not freely available. It is less than ideal that the decision regarding which, if any, parts of the evaluation are made public, is being made by the executive director alone. This is a task that I argue should be made by the nonprofit's board of directors, as representatives of the public-at-large.

Naturally, this research has limitations that should be acknowledged. First, the results are from only one city in the United States and may have limited or no generalizability to other locations. Second, the number of responses is considerably smaller in 2008 than in 1999, and the response rate is lower. It is unknown how this affects the comparisons made. Third, the relative difference in the percent of agencies receiving United Way funds between the 1999 and 2008 responses appears important and should be considered in future analysis. In addition, many of the more theoretically oriented questions are not yet addressed, pending qualitative data collection from agency directors. Still, the data seem to be yielding some tantalizing information that shows that United Way efforts to promote the use of evaluations are working.

## **Conclusion**

As an example of why public accountability is important, let me relate a cautionary tale. Recently in Fort Worth, Texas, a city very near Dallas, a large county hospital director and some of his close associates lost their jobs. While the reasons are many, it came to light that the director had buried the results of a consultant's study that was highly critical of his performance and the impact his decisions were having on patient care quality. When the report was finally seen by the local newspaper and publicized, the board members supposedly overseeing the operation of the public hospital claimed they had never been told about the report's existence, much less its negative content.

Accountability has to mean that stakeholders can judge for themselves if an agency is both efficient and efficacious. Organizations that arrange to have their Form 990s and/or other financial information easily accessible (as through Guidestar.org) are not only following the law but also acting in their own best interests. Knowing that the organization is open to scrutiny

provides just a bit more reason for the board and staff members to do the right thing, all the time, even if it is tempting to try to hide some bit of information or behavior.

If individuals are encouraged to live their lives so that they would not be ashamed to see every waking moment exposed on a reality TV show, how much more so should nonprofit human service organizations be admonished to operate in such a way that their every deed could be publicized on the evening news show? Individuals still have the expectation that, as long as they harm no one, some elements of their lives can remain private. Nonprofits, with the benefits they receive from society at large, and in the current political climate, should not have any such expectations, nor should they work to create any. On the contrary, they should be working to remove any traces of opaqueness.

While progress is being made, as shown in this research, it is incumbent on nonprofit leaders to be yet more open and accountable, working to keep their deeds accessible to all, encouraging each other to go the extra miles necessary to be fully accountable. Scholars and researchers need to maintain a focus on this topic, working together with those in the field to devise non-onerous ways of tracking results, and publicizing them appropriately. In this way, the premises and promises of accountability and one tool to achieve it, program evaluation, will be made manifest.

## References

- Bailey, M. (2005). Think “results,” not evaluation. *Public Manager*, 34(1), 8-10.
- Barman, E. (2002). Asserting difference: The strategic response of nonprofits to competition. *Social Forces*, 80(4), 1191-1222.
- Brown, L. & Troutt, E. (2004). A cooperative approach to accountability: Manitoba’s Family Violence Prevention Program. *International Journal of Public Administration*, 27(5), 309-330.
- Brown, L. & Troutt, E. (2007). Reporting does not equal accountability! The importance of funding characteristics and process on accountability. *International Journal of Public Administration*, 30(2), 208-225.
- Burnley, C., Mathews, C. & McKenzie, S. (2005). Devolution of services to children and families: The experience of NPOs in Nanimo, British Columbia, Canada. *Voluntas*, 16(1), 69-87.
- Carson, E. (2002). Public expectations and nonprofit sector realities: A growing divide with disastrous consequences *Nonprofit and Voluntary Sector Quarterly*, 31(3), 429-436.
- Chisholm, E. (1994). Evaluation: Where we are. *Evaluation Practice*, 15, 339-345.
- Cohen, R. (2006, December 8). Are tougher philanthropy laws needed? *CQ Researcher*, 16(43), 1025.
- Ebrahim, A. (2005). Accountability myopia: Losing sight of organizational learning. *Nonprofit and Voluntary Sector Quarterly*, 34(1), 56-87.
- Ferguson, A. (2005, March 24). Charity cases? *BRW*, 27(11), 44-54.
- Fischer, R. & Stelter, J. (2007). Testing faith: Improving the evidence base on faith-based human services. *Journal of Religion and Spirituality in Social Work*, 25(3/4), 105-118.
- Flack, T. & Ryan, C. (2005). Financial reporting by Australian nonprofit organizations: Dilemmas posed by government funders. *Australian Journal of Public Administration*, 64(3), 69-77.
- Flanigan, S. (2007). Paying for God’s work: A rights-based examination of faith-based NGOs in Romania. *Voluntas*, 18(2), 156-175.
- Frumkin, P. (2001). *Managing for outcomes: Milestone contracting in Oklahoma*. Arlington, VA: PricewaterhouseCoopers, The Endowment for the Business of Government.

- Hoefler, R. (2000). Accountability in action? Program evaluation in nonprofit human services agencies. *Nonprofit Management and Leadership*, 11(2), 167-177.
- Houchin, S. (2002). Holding ourselves accountable: Managing by outcomes in Girls, Incorporated. *Nonprofit and Voluntary Sector Quarterly*, 31(2), 271-277.
- Irvin, R. (2005). State regulation of nonprofit organizations: Accountability regardless of outcomes. *Nonprofit and Voluntary Sector Quarterly*, 34(), 161-178.
- Lee, M. (2004). Public reporting: A neglected aspect of nonprofit accountability. *Nonprofit Management and Leadership*, 15(2), 169-185.
- Martin, L., & Kettner, P. (1997). Performance measurement: The new accountability. *Administration in Social Work*, 21(1), 17-29.
- Marvel, M. & Marvel, H. (2007). Outsourcing oversight: A comparison of monitoring for in-house and contracted services. *Public Administration Review*, 67(3), 521-530.
- Mindel, C. & Hoefler, R. (2006). An evaluation of a family strengthening program for substance abuse offenders. *Journal of Social Service Research*, 32(4), 23-38.
- O'Reagan, K. & Oster, S. (2002). Does government funding alter nonprofit governance? Evidence from New York City nonprofit contractors. *Journal of Policy Analysis and Management*, 21(3), 359-379.
- Ospina, S., Diaz, W., & O'Sullivan, J. (2002). Negotiating accountability: Lessons from identity-based nonprofit organizations. *Nonprofit and Voluntary Sector Quarterly*, 31(1), 1-31.
- Roberts, N. (2002). Keeping public officials accountable through dialogue: Resolving the accountability paradox. *Public Administration Review*, 62(6), 658-699.
- Rossi, P., Lipsey, M., & Freeman, H. (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage.
- Stephenson, M. (2006). The Nature Conservancy, the press and accountability. *Nonprofit and Voluntary Sector Quarterly*, 35(3), 345-366.
- Tassie, B., Murray, V., & Cutt, J. (1998). Evaluating social service agencies: Fuzzy pictures of organizational effectiveness. *Voluntas*, 9, 59-79.
- Walden, G. (2006). Who's watching us now? The nonprofit sector and the new government by surveillance. *Nonprofit and Voluntary Sector Quarterly*, 35(4), 715-720.

Table 1: Mean attitudes about evaluation statements, 1999 and 2008

Statement	1999 (n = 91)	2008 (n = 27)	P value
<i>Broadly speaking, evaluations are useful</i>	6.16	6.26	.687
<i>An evaluation can help improve the program</i>	6.01	6.56	.035
<i>An evaluation can inform funders of good things the program does</i>	5.77	6.44	.015
<i>The cost of an evaluation is money well-spent</i>	5.30	5.33	.919
<b>Evaluations are usually biased or inaccurate</b>	4.99	2.37	.000
<b>The strengths and weaknesses of the program are already known</b>	4.45	3.48	.010
<b>Evaluations can hurt funding if the results are negative</b>	3.41	3.44	.919
<b>An evaluation is too much work</b>	2.92	2.93	.992
<b>Evaluation results are generally of no use to direct practice staff</b>	2.56	1.70	.007
<b>Evaluations merely state the obvious</b>	2.38	2.15	.424
<b>Evaluations are inherently more trouble than they are worth</b>	2.35	2.04	.251
<b>No evaluation is needed because we already know that the program is doing a good job</b>	2.33	1.85	.154
<b>If the results of an evaluation are negative, the evaluation is probably flawed</b>	2.13	1.89	.330

Possible responses are: 1 (Very much disagree); 2 (Fairly much disagree); 3 (Slightly disagree); 4 (Neither disagree nor agree); 5 (Slightly agree); 6 (Fairly much agree); and 7 (very much agree).

Note: statements in an *italic typeface* are positive towards evaluation; statements in a **bold typeface** are negative towards evaluation.