Evaluating Performance Measurement Systems in Nonprofit Agencies: The Program Accountability Quality Scale (PAQS)

DENNIS L. POOLE, JOAN NELSON, SHARON CARNAHAN, NANCY G. CHEPENIK, AND CHRISTINE TUBIAK

ABSTRACT

The drive for accountability in human services puts pressure on nonprofit agencies to develop performance measurement systems. But efforts to build capacity in this area have been hindered by the lack of instruments to evaluate the quality of proposed performance measurement systems. The Performance Accountability Quality Scale (PAQS) attempts to fill this gap. The instrument was field-tested on 191 program performance measurement systems developed by nonprofit agencies in Central Florida. Preliminary findings indicate that PAQS provides a structure for obtaining expert opinions based on a theory-driven model about the quality of a proposed measurement system in a not-for-profit agency. The instrument also is useful for assessing agency needs for technical assistance and for evaluating progress in the development of performance measurement systems. Further study is needed to test PAQS in other settings and to explore new areas of research in outcome evaluation.

INTRODUCTION

The drive for increased accountability in human services has put enormous pressure on nonprofit agencies to develop performance measurement systems. A great deal of the pressure comes from government, a chief funding source of services in the not-for-profit sector. The Government Performance and Results Act of 1993 (GPRA) requires all major programs funded by the federal government to measure progress toward goals, including

Dennis L. Poole, Ph.D. • School of Social Work, The University of Texas at Austin, 1925 San Jacinto Boulevard, Austin, TX 78712-1203; Tel.: (512) 232-5913; Fax: (512) 471-9600; E-mail: dennispoole@mail.utexas.edu.

American Journal of Evaluation, Vol. 21, No. 1, 2000, pp. 15–26.
ISSN: 1098-2140

outcomes (U.S. Government Accounting Office, 1996). State and local governments have followed suit, building performance measurement into their contracts with nonprofit agencies. Managed care companies require nonprofits to measure outcomes as part of their certification process for payment. National nonprofit organizations such as Big Brothers Big Sisters, Boys and Girls Club, and the Child Welfare League want local affiliates to prove that their programs make a difference in the lives of people (Hatry, 1997; Plantz, Greenway, & Hendricks, 1997).

But developing performance measurement systems is easier said than done. In the United States there are some 83,000 nonprofit agencies with annual incomes above $25,000 (Hodgkinson & Weitzman, 1996). Over the past twenty-five years, these agencies have been largely held accountable for outputs (i.e., number of units of service and number of people served) rather than outcomes (i.e., changes in participant knowledge, attitudes, values, skills, behavior, condition, or status). As a result, many of these agencies are having trouble today making the transition to "performance measurement," which includes outputs *and* outcomes (Newcomer, 1997).

In response, the United Way of America and other organizations are sponsoring efforts to build capacity in this area. Three important lessons have been learned thus far. First, it takes time to develop a sound performance measurement system. Eight steps have been identified, spanning three stages, including preparation, a trial run, and implementation (Hatry, van Houten, Plantz, & Greenway, 1996). Second, capacity building requires a substantial investment of resources. Extensive hands-on training and ongoing technical assistance are needed throughout the process (Plantz et al., 1997). And third, capacity building demands patience. Facilitation of organizational learning has been identified as the primary role of funders and evaluators during the developmental stages (Segal, 1997).

Despite progress, capacity-building efforts have been hindered by the lack of instruments to judge the quality of a performance measurement system (Lambur, 1993; Weiss, 1997). Without such tools, it is difficult to assess whether a proposed system will generate useful data on program outcomes. A good tool will identify the parts that are flawed and pinpoint where additional technical assistance is needed to improve the capacity of the system. In the field of program evaluation, these challenges fall under the domain of *evaluability assessment* (Rutman, 1980; Schmidt, Scanlon, & Bell, 1979; Wholey, 1979). This preevaluation step identifies problems that may hinder program performance, or may interfere with the generation of useful data. Evaluability assessment increases the likelihood that a proposed measurement system will deliver what it promises. And, if the design of the system is found to be unsound or seriously flawed, improvements can be made, avoiding the waste of staff time and agency resources.

We attempt to fill the gap in evaluability assessment for performance measurement systems and to stimulate further research in this area by designing, testing, and refining one instrument—the Program Accountability Quality Scale (PAQS). Findings to date indicate that PAQS provides a useful structure for obtaining expert opinions based on a theory-driven model about the quality of a proposed measurement system in a not-for-profit agency. PAQS also is useful in assessing agency needs for technical assistance and measuring progress in the development of a sound performance measurement system. In this article we provide background information on the development of PAQS at Heart of Florida United Way in Orlando, Florida. We also present preliminary findings on the reliability, structure, and utility of the instrument, and we suggest directions for future research.

DEVELOPMENT OF PAQS

Heart of Florida United Way (HFUW) is the largest nonprofit fundraising organization in Central Florida. Its annual budget of approximately $18 million supports nearly 180 human services programs through 76 member agencies. Fifty professional and support staff members are employed by the organization, and more than 500 volunteers are involved in planning, development, and fund distribution. The HFUW service delivery area spans Orange, Osceola, and Seminole Counties, commonly known as metropolitan Orlando.

The process that led to development of PAQS took four years. In 1995, HFUW began to search for ways to respond to national and local demands for increased accountability in human services. Following a series of discussions with the United Way of America and several local affiliates, HFUW in August 1996 hired a consultant, Dr. Jane Reisman, to infuse outcome evaluation into its fund distribution process. Over a 15-month period she conducted a series of meetings and workshops on performance measurement with HFUW and its local affiliates. Her efforts focused mainly on the development of logic models, along with key indicators and an evaluation plan. Logic models have proven helpful in cause-and-effect evaluations, and they have been used successfully to achieve consensus among stakeholders about program theory and desired outcomes (Wholey, 1979, 1983; Wong–Rieger & David, 1995). The logic model introduced by the consultant incorporates both *process evaluation* (resources, activities, and outputs) and *outcome evaluation* (outcomes and goals). It is a written tool that can be used to develop or identify a consistent thread of logic flowing through the resources, activities, outputs, outcomes, and goals of a program (Reisman, 1994; Reisman & Clegg, 1999).

The next step in the process required local affiliates to submit program logic models, along with indicators and an evaluation plan, in their funding application for fiscal year 1998/99. Soon thereafter, HFUW faced two major challenges. The first challenge was to judge the quality of the proposed systems efficiently, and then give consistent feedback to agency directors and program managers. The second challenge was to determine whether HFUW's investments in capacity building were paying off, and to identify what additional investments in training and technical assistance would be needed to improve the quality of these systems. Because no tools were available, we formed a research team that developed an instrument that would help HFUW meet these two challenges.

STUDY

The original field-tested version of PAQS consisted of 34 items. Based on current theoretical knowledge about performance measurement (Hatry et al., 1996; Rutman, 1980; Wholey, 1979), we hypothesized that a sound performance measurement system would consist of seven domains. Thus, items in PAQS were constructed to represent seven subscales:

- Resources—program ingredients (e.g., funds, staff, community support, participants);
- Activities—methods used to accomplish program goals (e.g., classes, counseling, training);

- Outputs—units produced by a program (e.g., number and type of clients served, number of policies developed, number of events planned);
- Outcomes—short and immediate indicators of progress toward goals (e.g., improved school-related behaviors, increased parental knowledge of child development, improved family functioning);
- Goals—long-term desired program effects (e.g., resilient community, economic self-sufficiency, violence prevention);
- Indicators—specific and observable terms to measure whether a program has achieved an intended outcome (e.g., grades, attendance, discipline reports, scores on family functioning scale, scores on knowledge of child development test); and
- Evaluation Plan—a systematic method to generate reliable and valid data to measure progress toward outcomes (e.g., measurement tools, data collection procedures, sampling strategy) (Reisman, 1994).

Two evaluators conducted assessments on 191 program performance measurement systems submitted by 78 nonprofit agencies seeking funds from Heart of Florida United Way, Orange County, or City of Orlando Citizens' Review Panel in 1998. Each system was described by its respective agency on a three-page form. The first page required agencies to identify elements of the program logic model (resources, activities, outputs, outcomes, and goals), the second to list key indicators of each outcome, and the third to delineate the evaluation plan (measurement tools, data collection procedures, and sampling strategy).

The evaluators, who had an average of 15 years' experience in program evaluation, were familiar with the training materials developed by Reisman. HFUW staff randomly assigned half of the proposed performance measurement systems—that is, cases—to each evaluator. To assess interrater reliability, a random sample of eight cases was assigned to each evaluator. The evaluators then independently rated the cases on each of the 34 items in the original PAQS instrument, without knowing which cases served as the reliability check.

A similar procedure was followed a year later, when a third evaluator (with similar expertise) was added to the research team. Using a revised 21-item version of PAQS (Appendix), each of the three evaluators independently rated one-third of the proposed performance measurement systems submitted to HFUW in 1999. Reporting practices for the agencies did not vary from the previous year, and virtually the same programs were represented in the study population. HFUW staff randomly assigned one-third of the proposed systems to each of the three evaluators. To assess interrater reliability, a random sample of 20 cases was assigned to each evaluator. The evaluators then independently rated the cases on each of the 21 items of the revised instrument, blind as to which cases served as the reliability check. Rating and scoring a case usually took about 25 minutes.

FINDINGS

The sample of 191 program performance measurement systems was distributed across HFUW's seven "focus care areas" as follows: Helping Children Learn and Grow (8.4%); Guiding Our Youth Toward Success (19.9%); Assisting and Strengthening Individuals and Families (13.9%); Enabling Seniors to Remain Independent and Active (12.0%); Helping Individuals with Disabilities Develop Their Potential (10.8%); Providing People with Emergency Assistance and Shelter (19.3%); and Improving Local Health Needs (15.7%). The

median total program budget was $241,736, with a range of $8,000 to $24,065,187. Total agency budgets ranged from $15,149 to $30,195,939.

Reliability

We first analyzed inter-item correlations and item-total correlations among the responses from the sample. Using the domain model of sampling (Nunnally & Bernstein, 1994), we eliminated 13 items that were not strongly associated with the total set of 34 items in the original version of PAQS. This reduced the length of the scale from 34 to 21 items. The corrected item-total correlations for the retained items ranged from .23 to .65.

To measure the consistency of responses among the 21 items, Cronbach alphas were computed for PAQS and its subscales. The overall reliability estimate for the revised scale was alpha = .84, indicating good internal consistency. The reliability coefficients for the seven subscales were as follows: Resources (.86), Activities (.72), Outputs (.74), Outcomes (.76), Goals (.64), Indicators (.78), and Evaluation Plan (.84).

We also examined interrater reliability. As noted above, a random sample of eight cases was assigned to the two evaluators during the first year of the project. Given this small number of cases, we calculated the percent of agreement on item responses (Rubin & Babbie, 1997). The overall average rate of agreement among the 21 items in the revised scale was high (83%), despite the small sample. During the second year of the project, interrater reliability was analyzed again, this time with a random sample of 20 cases randomly assigned to each of the three evaluators. The reliability estimates were high, with correlation coefficients across each pair of raters on overall PAQS scores ranging from .81 to .92.

Subscale Structure

We then used the multiple groups centroid method to examine the structure of PAQS. This method identifies underlying properties (domains) that unite certain items on a scale (Kerlinger, 1986). The analysis was conducted on the revised version of each domain in PAQS. In other words, those items deleted based on reliability analysis were not included in the factor analysis. The number of factors was fixed at seven, equal to the number of subscales that we hypothesized in PAQS. As a general rule, item correlations with factors are considered moderately high when they fall around .60 (Nunnally & Bernstein, 1994).

Results of the factor analysis reveal that items loaded well on the domains with which they were intended (Table 1). This is indicated by the factor loadings for each item, italicized in bold under each item's respective domain. Thus, there is evidence that the structure of PAQS corresponds to the conceptual framework that underlined its development. No established empirical benchmarks were available to examine construct validity, which requires evidence of convergence and divergence.

Utility

To examine the utility of PAQS, we considered its usefulness as a tool to judge the quality of a proposed system, assess technical assistance needs, measure system-wide progress toward performance measurement, and provide feedback to agency directors and program managers. Table 2 presents PAQS scores on a sample of 154 cases of proposed measurement systems submitted to HFUW in 1998. The scores are listed from high (84) to

TABLE 1.
Factor Analysis of PAQS: Factors and Loadings

	Factors						
Item	*Resource*	*Activity*	*Outputs*	*Outcomes*	*Goal*	*Indicators*	*EvalPlan*
Resource 1	***.883***	.264	.266	.021	.215	.124	.081
Resource 2	***.939***	.300	.170	.200	.209	.068	.200
Resource 3	***.872***	.268	.152	.099	.167	.051	.187
Activity 4	.239	***.877***	.218	.305	.107	.246	.284
Activity 5	.307	***.900***	.375	.186	.254	.311	.338
Output 6	.126	.224	***.750***	.175	.030	.080	.165
Output 7	.198	.332	***.901***	.172	.110	.218	.222
Output 8	.185	.287	***.936***	.119	.121	.186	.198
Outcome 9	.131	.184	.013	***.618***	.040	.144	.116
Outcome 10	.044	.293	.237	***.905***	.079	.380	.345
Outcome 11	.134	.193	.172	***.914***	.026	.329	.281
Goal 12	.074	.160	.080	.034	***.811***	.193	.068
Goal 13	.282	.188	.096	.027	***.875***	.071	.098
Indicator 14	.135	.359	.217	.215	.053	***.691***	.426
Indicator 15	.002	.090	.113	.170	.136	***.800***	.322
Indicator 16	.022	.218	.036	.386	.171	***.853***	.456
Indicator 17	.143	.316	.260	.354	.137	***.800***	.493
Evaluation Plan 18	.308	.172	.105	.191	.093	.353	***.734***
Evaluation Plan 19	.082	.307	.219	.199	.076	.322	***.793***
Evaluation Plan 20	.058	.336	.168	.295	.132	.475	***.870***
Evaluation Plan 21	.189	.302	.239	.269	.084	.521	***.910***

low (21), with the upper half of the scores divided into intervals of 10%. Results indicate that only 4 (2.6%) of the proposed systems scored between 78 and 84, that is, they received at least 90% of the total possible points on PAQS. Thirty-eight (24.7%) scored in the 80% range (71 to 77 points), and 46 (29.8%) scored in the 70% range (64-70 points). In contrast, 66 (42%) of the proposed systems received ratings at or below 60% of total points on the scale.

There was no specific point in the distribution of PAQS scores where HFUW deemed a proposed system "sound" or "unsound," "acceptable" or "unacceptable." Cutoff scores usually cannot be set scientifically because in most instances they are "a matter of professional judgment," based on values, practical considerations, and organizational goals (Cascio, Alexander, & Barrett, 1988). Still, PAQS helped HFUW rank scorers from high to low and provided baseline data on the quality of the performance measurement systems at this stage of the project. The good news was that more than one-half of the proposed systems scored 70% or more of total possible scores on the scale, providing some evidence that HFUW's capacity-building efforts were "paying off." On the other hand, the broad distribution of scores revealed that capacity building takes time; additional investments would be needed to improve scores throughout the HFUW system.

More importantly, PAQS helped HFUW pinpoint where additional investments in technical assistance were needed. The sample was divided into two groups—those that

TABLE 2.
PAQS Scores, Total Sample By Frequency and Percentage, 1998

Score	*Frequency*	*Percent*	*Cumulative Percent*	*Percent of Total*
84	0	—	—	
83	0	—	—	
82	0	—	—	
81	0	—	—	90%
80	0	—	—	
79	0	—	—	
78	4	2.6	2.6	
77	7	4.5	7.1	
76	2	1.3	8.1	
75	3	1.9	10.4	
74	4	2.6	13.0	80%
73	5	3.2	16.2	
72	14	9.1	25.3	
71	3	1.9	27.3	
70	4	2.6	29.9	
69	5	3.2	33.1	
68	6	3.9	37.0	
67	10	6.5	43.5	70%
66	9	5.8	49.4	
65	7	4.5	53.9	
64	5	3.3	57.1	
63	2	1.3	58.4	
62	10	6.5	64.9	
61	5	3.2	68.2	
60	6	3.9	72.1	60%
59	2	1.3	73.4	
58	3	1.9	75.3	
57	4	2.6	77.9	
56	5	3.2	81.2	
55	3	1.9	83.1	
54	4	2.6	85.7	
53	4	2.6	88.3	
52	1	.6	89.0	
51	5	3.2	92.2	
50	3	1.9	94.2	50%
49	2	1.3	95.5	or Lower
48	2	1.3	96.8	
47	2	1.3	98.1	
46	0	—	—	
45	2	1.3	99.4	
44	1	.6	100.0	
21–43	0	—		

scored a 3 or 4 on an item (in the Very Good to Good range) and those that scored a 1 or 2 (in the Fair to Poor range). As shown in Table 3, the large majority of the proposed systems scored in the Very Good to Good range on the items in Resources, Activities, and Goals. But scores on many items in the other five subscales were much lower. For example, 90.5% of

TABLE 3.
PAQS Subscale/Item Scores by Percentage, 1998

Subscale/Items	*Very Good to Good Performance Range*	*Fair to Poor Performance Range*
	%	%
Resources		
1. Identification of resources	78.5	21.5
2. Comprehensiveness	81.6	18.4
3. Match the type of program	90.5	9.5
Activities		
4. Logical link to outputs	91.1	8.9
5. Sufficient activities to achieve outcomes	89.9	10.1
Outputs		
6. Number of participants	69.0	31.0
7. Number of events or processes	71.5	28.5
8. Time frame	69.0	31.0
Outcomes		
9. Logical link to goals	90.5	9.5
10. Change statements	60.8	39.2
11. Outcomes rather than activities or outputs	63.3	36.7
Goals		
12. Intended effect on need and population	93.6	6.4
13. Description of broad community impact	87.3	13.7
Indicators		
14. Specific and measurable terms	72.2	27.8
15. Valid measures of outcomes	60.3	39.7
16. Efficient measures	68.4	31.6
17. Important to changes in need of measurement	75.9	24.1
Evaluation Plan		
18. Data collection method	65.2	34.8
19. Resources for implementation	89.8	10.2
20. Efficient measurement of progress toward outcomes	72.6	27.4
21. Realistic plan	65.0	35.0

the proposed systems logically linked outcomes to goals (item #9), but only 60.8% had outcomes that were written as change statements (item #10), and only 60.3% had indicators that were valid measures of the outcomes (item #15). Based on the data reported in Table 3, HFUW would need to invest additional resources in technical assistance to improve system-wide scores on PAQS, especially in the components dealing with Outputs, Outcomes, Indicators, and Evaluation Plans.

Recognizing that HFUW did not have enough resources to provide intensive technical assistance in all areas, staff decided to focus their efforts in the second year on improving scores in the logic models, especially Outputs and Outcomes. This strategy apparently worked. As Table 4 shows, when the measurement systems were reevaluated in 1999, the number that scored 90% or more (78 to 84 points) on PAQS increased dramatically, from 4 (2.6%) to 35 (22.2%); the number scoring at or above 80% jumped from 38 (24.7%) to 46 (29.1%). In contrast, scores at or below 70% dropped from 112 (72.7%) to 77 (48.7). Further, paired sample *t*-tests (Table 5) revealed that total PAQS scores improved significantly from

TABLE 4.
Comparison of PAQS Scores, 1998 and 1999

	1998		*1999*		
Score (Range)	*#*	*%*	*#*	*%*	*Change*
78–84 (90% range)	4	2.6	35	22.2	+31
71–77 (80% range)	38	24.7	46	29.1	+8
64–70 (70% range)	46	29.9	34	21.5	−12
≥63 (60% or lower)	66	42.8	43	27.2	−23

1998 to 1999. Statistically significant improvements in subscale scores were found for Activities, Outputs, and Outcomes. As expected, scores did not improve significantly for Goals and Resources (which were relatively high during the first year of the project), nor did they improve for Indicators and Evaluation Plan (which were not a primary focus of technical assistance activities during the second year). Thus, PAQS helped HFUW generate baseline data on the quality of the proposed systems, pinpoint technical assistance needs, and assess whether their second-year investments in capacity building had paid off.

PAQS provided a mechanism for HFUW to evaluate proposed measurement systems efficiently, then give consistent feedback to agency directors and program managers. The contribution of PAQS in these areas is critical. Funders often go through the motions of outcome evaluation, but do not hold agencies accountable for the quality of their systems because they lack either confidence in their ability to evaluate these systems consistently or the resources to do it. PAQS provided a structure for HFUW to obtain expert opinion on the quality of a proposed system at an average cost of $20 an assessment. The instrument also helped staff give consistent—and detailed—feedback on the strengths and weaknesses of each proposal as well as its level of quality in relation to other proposals. In this way, the instrument clarified expectations for both HFUW and agencies.

TABLE 5.
Paired Samples T-Test of PAQS Scores, 1998 and 1999

	Paired Differences		
Scale/Subscale	*t*	*df*	*Sign. (2 tailed)*
Total Scale	4.95	156	.001
Resources	−.18	157	.858
Activities	11.32	157	.001
Outputs	5.85	157	.001
Outcomes	5.84	157	.001
Goals	−.75	156	.457
Indicators	−1.32	155	.188
Evaluation Plan	−1.04	156	.302

CONCLUSION

The Performance Accountability Quality Scale (PAQS) is a 21-item instrument with supportive evidence concerning its reliability, structure, and utility. It provides a tool for obtaining expert opinions based on a theory-driven model about the quality of a proposed measurement system in not-for-profit agencies. PAQS also is useful in identifying agency needs for technical assistance, in assessing progress in the development of performance measurement systems, and in giving efficient and consistent feedback to agency directors and program managers.

Such inferences must be constrained by four considerations. First, data were drawn from a sample of United Way-supported nonprofit programs in one geographical region. Further study is needed to test whether the findings can be generalized to programs in other settings. Second, the construct validity of the instrument has not been assessed. An empirical benchmark—based on actual program outcomes—needs to be identified or developed for this purpose. Third, more research, involving a larger number of cases and evaluators with different levels of experience, should be conducted to test for interrater reliability. This will be especially important when funding staff and lay volunteers join professional evaluators in judging the quality of performance measurement systems. Finally, the instrument only includes criteria drawn from one theoretical model of performance measurement. Other criteria have been identified by experts in outcome evaluation, such as program theory, intermediate goals and objectives, mechanisms through which a program is expected to have its effects, and the integrity of an agency's information system to measure outcomes (Reisman & Clegg, 1999; U. S. Government Accounting Office, 1999). Consideration should be given to modifying PAQS to add these criteria.

To those who study nonprofit agencies, PAQS might be useful for other research purposes. For one, little is known about predictors of success in the development of performance measurement systems. Do variables such as agency size, program type, revenue source, and budget size account for variation in readiness scores? Or do other factors such as agency culture, leadership, technology, and staff make the difference? The instrument also could be used to examine questions related to power and perceived risk. Do agency executives believe that their programs will actually be held accountable for outcomes? Or do they believe that power and political expediency will drive future funding decisions in the nonprofit sector? The development of tools such as PAQS enables us to explore systematically how these factors influence the development of performance measurement systems in nonprofit agencies.

REFERENCES

Cascio, W. E., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology, 41*, 1–24.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart & Winston.

Hatry, H. (1997). Outcome measurement and social services: Public and private sector perspectives. In E. J. Mullen & J. L. Magnabosco (Eds.), *Outcomes measurement in the human services* (pp. 3–19). Washington, DC: NASW Press.

Hatry, H., van Houten, T., Plantz, M. C., & Greenway, M. T. (1996). *Measuring program outcomes: A practical approach.* Alexandria, VA: United Way of America.

Hodgkinson, V., & Weitzman, M. (1996). *Nonprofit almanac 1996–1997.* San Francisco: Jossey-Bass.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31–36.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Reinhart & Winston.

Lambur, M. (1993, November). *Can evaluability assessment be usefully shortened for use in small independent organizations?* Paper presented at the American Evaluation Association Annual Meeting, Dallas, Texas.

Newcomer, K. E. (1997). Using performance measurement to improve programs. In K. E. Newcomer (Ed.), *Using performance measurement to improve public and nonprofit programs* (pp. 5–13). San Francisco: Jossey-Bass.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Plantz, M. C., Greenway, M. T., & Hendricks, M. (1997). Outcome measurement: Showing results in the nonprofit sector. In K. E. Newcomer (Ed.), *Using performance measurement to improve public and nonprofit programs* (pp. 15–30). San Francisco: Jossey-Bass.

Reisman, J. (1994). *A field guide to outcome-based program evaluation.* Seattle, WA: The Evaluation Forum, Organizational Research Services, Clegg & Associates.

Reisman, J., & Clegg, J. (1999). *Outcomes for success 2000.* Seattle, WA: The Evaluation Forum, Organizational Research Services, Clegg & Associates.

Rubin, A., & Babbie, E. (1997). *Research methods for social work* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Rutman, L. (1980). *Planning useful evaluations: Evaluability assessment.* Beverly Hills: Sage.

Schmidt, R. E., Scanlon, J. W., & Bell, J. B. (1979). *Evaluability assessment: Making public programs work better.* Rockville, MD: U.S. Department of Health, Education and Welfare, Project Share, Human Services Monograph 14.

Segal, S. P. (1997). Outcomes measurement systems in mental health: A program perspective. In E. J. Mullen & J. L. Magnabosco (Eds.), *Outcomes measurement in the human services* (pp. 149–159). Washington, DC: NASW Press.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

U. S. Government Accounting Office (1996, June). *Effectively implementing the Government Performance and Results Act* (GAO/GGD-96-118). Washington, DC.

U. S. Government Accounting Office (1999, February). *Agency performance plans: Examples of practices that can improve usefulness to decision makers* (GAO/GGD/AIMD-99-69). Washington, DC.

Weiss, H. B. (1997). Results-based accountability for child and family services. In E. J. Mullen & J. L. Magnabosco (Eds.), *Outcomes measurement in the human services* (pp. 173–180). Washington, DC: NASW Press.

Wholey, J. S. (1979). *Evaluation: Promise and performance.* Washington, DC: The Urban Institute.

Wholey, J. S. (1983). *Evaluation and effective public management.* Boston: Little, Brown.

Wong-Rieger, D., & David, L. (1995). Using program logic models to plan and evaluate education and prevention programs. In A. J. Love (Ed.), *Evaluation methods sourcebook II* (pp. 120–135). Ottawa, Ontario: Canadian Evaluation Society.

Appendix
Program Accountability Quality Scale (PAQS)©

Resources
Most areas of resources are addressed.
The resources seem comprehensive.
The resources seem to match this type of program.

Activities
The activities logically link to the outputs listed.
There are sufficient activities to achieve the outcomes.

Outputs
The numbers of participants are identified for each activity.
The numbers of events/processes are listed.
Time frames are given for outputs.

Outcomes
The outcomes logically link to the goal(s).
The outcomes are written as change statements.
The outcomes are truly outcomes rather than activities or outputs.

Goals
The program goals indicate the intended effect of the program on the need and population.
The program goals describe the broad community impact.

Indicators
The indicators are stated in specific and measurable terms.
The indicators are valid measures of the outcomes.
The indicators will efficiently measure progress toward achievement of the outcomes.
The indicators are important to the changes program planners want to measure.

Evaluation Plan
The data collection method will generate reliable information.
The evaluation plan can be implemented with available resources.
The evaluation plan is designed to measure progress toward outcomes in an efficient manner.
The evaluation plan is realistic.

NOTE: Rating response options for each item include (4) Strongly Agree/Very Good Performance—only minor revisions needed, if any; (3) Agree/Good Performance—some strengths, some areas need revision; (2) Disagree/Fair Performance—few strengths, major revisions required; and (1) Strongly Disagree/Poor Performance—lacking strengths, insufficient information provided.

Copies of the Program Accountability Scale (PAQS) can be obtained at Heart of Florida United Way, Dr. Nelson Ying Center, 1940 Traylor Boulevard, Orlando, Florida 32804 (407-835-0900).