Common Evaluation Tools Across Multi-State Programs: A Study of Parenting Education and Youth Engagement Programs in Children, Youth, and Families At-Risk

Abstract

Community-based education programs must demonstrate effectiveness to various funding sources. The pilot study reported here (funded by CYFAR, NIFA, USDA award #2008-41520-04810) had the goal of determining if state level programs with varied curriculum could use a common evaluation tool to demonstrate efficacy. Results in parenting and youth engagement indicated that with effort to select valid and reliable measures, it is possible to use common measure across curricula. Lessons learned including evaluating goodness of fit are discussed in regards to the process of conducting common measures evaluations.

Pamela B. Payne
Assistant Professor
Jerry and Vicki Moyes
College of Education
Weber State
University
Ogden, Utah
pamelapayne@weber.
edu

Daniel A. McDonald
Area Associate Agent
Pima County
Cooperative
Extension
The University of
Arizona
Tucson, Arizona
mcdonald@cals.arizon
a.edu

Introduction

In the last decade, it has become more difficult for community-based programs to receive funding at high levels in order to develop and sustain programming (Mancini & Marek, 2004). This challenge has become particularly salient for federally funded programs that use different evaluation measures, making the comparison of program outcomes across the country challenging, if not impossible. This was the case with community programs funded through the Children, Youth, and Families At-Risk (CYFAR) initiative at the United States Department of Agriculture (USDA)/National Institute of Food and Agriculture (NIFA). Program leaders needed to demonstrate the efficacy of the community programs funded through CYFAR; however, very little could be reported about collective program outcomes (Roucan-Kane, 2008; Adler-Baeder, Kerpelman, Griffin, & Schramm, 2010).

To remedy this situation a task force was created in 2008 to assess the feasibility of having all CYFAR Sustainable Community Projects (SCP) use common evaluation tools based on program content areas across the United States. This article discusses how the use of a common evaluation instrument was established in two specific program areas: parenting and youth citizenship; the lessons learned as common measures were implemented; and questions organizations may ask before using common

measures to evaluate their community programs.

Developing a Common Measures Plan

Initially, a performance monitoring framework was implemented to track program progress across global indicators identified for each outcome cluster of CYFAR programs. The initial clusters were based on the types of programs funded by CYFAR, including: youth citizenship, parenting, healthy lifestyles, workforce preparation, and literacy and communication. There were inherent challenges and limitations in simply assessing these global indicators without also using common instruments to measure outcomes in each area. Each SCP used their own evaluation tool, which made it difficult to reach any conclusions about the overall effectiveness of programs in any one domain or outcome cluster. In response to this situation, a pilot study was initiated with the goal of exploring the feasibility of using common evaluation tools in various outcome clusters to enable CYFAR to demonstrate effectiveness across the breadth of programs being funded.

Two outcome clusters, parenting and youth citizenship, were selected as foci for this pilot program. Sites that received funding for these clusters in 2009 were invited to participate in the pilot study. The goal of the task force was to solicit 10 sites to participate, five from each outcome area. Out of 13 potential sites, seven state SCP sites were successfully recruited to participate in this pilot program: Alaska (youth citizenship and parenting), Arizona (youth citizenship and parenting), Iowa (parenting), Kentucky (youth citizenship), New York (youth citizenship), Tennessee (parenting), and Washington (parenting). Some state SCP sites had both parenting and youth citizenship programs that participated in the common measure pilot study (five parenting and four youth citizenship). The remaining six sites elected not to participate for a variety of reasons, including not having ongoing programs during the time period of evaluation. Data collection began in the winter of 2010 and continued through July 2011.

A uniform instrument and methodology were provided to each participating SCP. An instrument designed to measure parenting program outcomes and a data collection protocol were provided to sites with parenting programs. Another instrument designed to measure outcomes associated with youth citizenship and a corresponding data collection protocol were provided to programs involved in youth citizenship. The project did not assume that the uniform instrument provided would necessarily be the only evaluation instrument used. The decision about including additional evaluation measures or approaches was left up to the discretion of each participating state. Pre-post data were used unless there was an existing cohort, and no new participants were expected to be recruited during the period of the pilot study, in which case a retrospective method was employed. An online survey was developed; however, sites could use paper and pencil surveys if data entry was performed by site staff and forwarded to the lead team.

Instrument Identification & Selection

A literature review was conducted of evaluation tools in the areas of parenting and youth citizenship to determine which instruments to use for the common measures pilot study. Measures were screened based on length, content, and psychometric properties. In consultation with pilot sites it was determined that the domain of parenting would focus on early childhood through adolescence and parents' reports of involvement and communication with their children. The domain of youth citizenship would focus on

areas such as youth voice, engagement, youth-adult partnerships, and leadership skills.

The principal investigator and project coordinator of the pilot study worked closely with task force members, project directors, and evaluators from pilot sites to vet instruments and design an appropriate data collection protocol. The process involved an in-person meeting during the CYFAR conference in 2009 and several subsequent conference calls and group email correspondence to finalize the selection of instruments to be used. Three primary criteria were identified in the selection of an instrument:

- Measures were to be short (approximately 10-20 items), concise, and easy to read and understand (translation of measures into Spanish was identified as a possibility);

- Measures should reflect the content areas and theoretical foundations of the programs studied; and

- Measures were to be well established, good reliability properties, and validated with populations served by the CYFAR projects.

For the parenting cluster, a combination of two existing scales was used as the common measure. The Intervention Targeting Parent Behavior scale (Spoth, Redmond, Haggerty, & Ward, 1995) is a 13-item scale that has been used in Native American, Latino, and African-American populations. This measure had good reliability established previously ranging from $\alpha = .68$ to $\alpha = .87$ (Spoth, et al., 1995).Three items on addiction and substance use behavior were eliminated from the common measure. The Parenting Pre-Education Survey ((Cornell Cooperative Extension Parent Education and Parenting Education Program Work Team (PEPWT)) is a 16-item measure focusing on assets rather than interventions, and was selected for its easy implementation and plain language (Baker & Mott, 1989; Cornell Cooperative Extension, 2009). The PEPWT also had good established reliability ranging from $\alpha = .50$ to $\alpha = .90$ (AZREACH, 2014).

For the youth citizenship cluster, a measure developed at the University of California Berkeley was selected as it met the criteria for inclusion and gained the greatest consensus for its use from the task force members and SCP site representatives. The UC Berkeley Civic Responsibility Scale (Furco, Muller, & Ammon, 1998) has 10 items that tap into the domains of civic responsibility in school-age and early adolescent youth. This measure had good established reliability of $\alpha = .84$ in previous studies (Santos, 1999).

All measures were offered in a pre/post format or retrospective format to accommodate variation in program length, cycle, and frequency. Given the cycle for some programs (e.g., programs were ongoing at the time data collection began), a pre-post format was not conducive to data collection, and in those cases a retrospective format was used. Measures were also made available in both paper/pencil format or online (via Survey Monkey and a website). Approximately half the sites implemented a pre/post format, while the other half used the retrospective format. Although the online formats were available, all sites elected to use a pencil/paper method and conduct data entry onsite.

Because this article is focused on the methodology associated with using the common measures in evaluation, information on specific outcome measures and results can be found in other articles related to the study reported here (Payne & McDonald, 2012; Payne & McDonald, 2014).

Lessons Learned and Challenges for Using Common Measures

Throughout the pilot study, participating SCPs were asked to comment on the challenges, advantages, disadvantages, and concerns they had with using common evaluation instruments. Participating SCPs provided feedback both formally and informally in quantitative and qualitative formats, using a Goodness of Fit survey, discussions during meetings, and panel presentations at conferences.

The pilot study was to explore the feasibility of using common measures for program evaluation within the CYFAR program. Therefore, the pilot study process incorporated opportunities for reflection on the process itself through a "Goodness of Fit" survey sent to site administrators. Lessons learned early on included:

- Soliciting input from participating sites early in the process regarding measures and method of data collection helped to ensure cooperation from sites;

- Explaining the value of using common measures and the benefits to sites helped to secure buy-in from participating sites; and

- Understanding current data collection capabilities of sites helped to tailor the pilot study method to meet the needs of participating sites.

Table 1 lists some of the primary challenges encountered and solutions employed during the pilot study.

Table 1.
Challenges and Solutions to Common Measures

Challenge	Solution Used in Current Pilot Study
Program delivery differs in terms of duration and intensity.	Measurement instruments included items pertaining to "dosage."
Programs cohorts may be continuous or changing.	A retrospective/reflective instrument was designed in addition to the pre/post instrument.
Some sites may have well established evaluations already underway.	The common measure can augment existing evaluations and provide comparative results.
Common measures may be redundant with existing measures.	Alterations were permitted so as to not repeat items or scales.
Institutional Review Board (IRB) approval might get complicated.	IRB approval was received from the institution conducting the pilot study. Each pilot site was responsible for individual site approval.

Not every challenge could be adequately resolved to the satisfaction of all sites. While every effort was made to identify the most relevant measures available, it was not possible to find one single measure

that satisfied the needs of every site. To assess the fit between the common measure and the program content an internal survey was administered to sites in the pilot study.

The Goodness of Fit survey was designed to elicit feedback on the appropriateness of the measures selected. It is crucial to involve evaluators in creating an expectation and understanding of the process (Boyd, 2009). During the selection process, SCP representatives voiced concern over some of the items included in the common measure evaluation instruments that were not necessarily covered by their programs. For instance, the parenting survey asks how often the parent reads to his or her child, or how often they discuss drug or alcohol use. Clearly these items were appropriate for some parenting programs and not for others. Similar issues arose when considering instruments for youth citizenship programs. These questions were used at all sites even if not directly related to specific program content to maintain the integrity of the common measure while not sacrificing reliability (Payne & McDonald, 2012, Payne & McDonald, 2014). SCP site representatives participating in the pilot study were encouraged to complete the Goodness of Fit survey. For each item contained in the parenting and youth citizenship surveys, respectively, program staff (N = 10) were asked to assess the relevance of each item to the topics covered in their CYFAR program. Possible responses included: Completely Relevant, Somewhat Relevant, Slightly Relevant, Not Relevant, and Not Sure. A summary of item relevancy results is shown in Table 2 across both parenting and youth citizenship programs.

Table 2.

Goodness of Fit Results

Item Relevance	% of 22 items relevant
Completely Relevant	66%
Somewhat Relevant	27%
Slightly Relevant	5%
Not Relevant	<2%
Not Sure	<1%
N = 10	

During a facilitated discussion of common measures, participating pilot study sites along with program directors and evaluation professionals from other CYFAR-funded projects were asked to identify some of the advantages of using common measures. A compilation of their responses is outlined in Table 3.

Table 3.

Advantages to Using Common Measures

Broad Implication	Program Development	Measurement
Better able to share value and effectiveness of program with clientele and stakeholders	Helps build more effective programs	Availability of vetted measures
Better able to market benefits	Enables sharing	Potential cost savings from

of program	of information across programs	reduced evaluation efforts locating or developing measures
Ability to spend more time on project because evaluation instruments and methods are set	Improves understanding of program content/outcomes	Common measures may provide some consistency in programming or help set standards
Better able to foster sustainability of program		
Ability to compare results with similar projects		
Ability to build local capacity for evaluation		
Provides direction and structure to sites struggling with evaluation goals		
Helps contribute to regional and nation-wide outcomes		

Discussion and Implications

The goal of the Common Measure Pilot Study was to assess the extent to which CYFAR SCP sites were willing and able to use common evaluation instruments and to obtain data from CYFAR SCP participants to be able to assess the collective impact of programs across the country.

Seven state CYFAR programs (five parenting & four youth citizenship) were represented in the pilot study. Two states had both parenting and youth citizenship programs and therefore collected data from both groups. 732 participants completed surveys for the evaluation on the feasibility of the common measure process. The majority of respondents from parenting programs were Hispanic (55%) women (70%). The majority of respondents from the youth citizenship programs were White (55%) females (53%). Having sites from Alaska to New York, and from Tennessee to Arizona, provided a good cross-section of CYFAR SCP program participants.

Of the 13 sites recruited, six CYFAR SCP sites decided not to participate in the pilot study. Seven sites did participate, and, of those, all seven elected to use paper and pencil surveys, although online surveys were available. Pilot sites were compensated for the additional expenses related to data collection, which may account for a site's willingness to participate. Participating as a pilot site required arranging sub-contracts between institutions and complying with Institutional Review Board requirements from the University of Arizona, in addition to the requirements of the participating university. Therefore, participation in the pilot study presented an additional administrative burden that would not be encountered if use of common instruments became a routine requirement of CYFAR funding. That being said, most sites were eager to incorporate the identified common instruments into their data collection

protocol.

Criteria for selecting the common measures were that they be validated and reliable measures. An analysis of the reliability of these measures shows that they continued to possess sound psychometric properties when used with CYFAR SCP populations. While two of the sub-scales, Anger Management (parenting survey) and Connection to Community (youth citizenship survey), showed a slightly lower Cronbach Alpha on the post-test sample (0.61 and 0.51, respectively), all the other sub-scales were comparable to psychometrics reported in previous studies (Furco, Muller, & Ammons, 1998; Spoth, et al., 1995). Furthermore, results from the goodness of fit examination indicates that most of the items (66%) included in the common measure surveys were deemed completely relevant to the content of the program examined.

Using common instruments across programs enables CYFAR to now speak with authority about its collective impact on outcomes related to participants' skills, knowledge, behaviors, and attitudes within the population studied. For both parenting and youth citizenship, significant increases were detected from pre-test to post-test, with large effect sizes measured for each sub-scale (Payne & McDonald, 2012; Payne & McDonald, 2014). The degree to which increases occurred did depend on the amount of intervention received by participants (Payne & McDonald, 2012; Payne & McDonald, 2014).

Overall the pilot study showed that CYFAR SCP sites were willing to participate in an evaluation that used common measurement instruments, provided those measures were valid, reliable, and did not place an unreasonable burden on the participant to complete (surveys that are short and to the point). It appears that using common measures is a viable and realistic way to evaluate the effectiveness of CYFAR SCP programs across the country in various domains. The use of common evaluation tools can also be beneficial for Extension in general and other large organizations (e.g., YMCA, Boys & Girls Clubs) that operate in various locations across the country and need to demonstrate the effectiveness of the program as a whole to funding sources. That being said, there are still a number of outstanding questions organizations may consider when deciding whether or not to use common measures.

- Will the results of common measures be used to decide funding in the future?

- Will comparisons be made that make some sites feels less adequate? Will results lead to punitive actions?

- Will language and cultural differences and literacy levels be taken into account when deciding on common measures?

- Will flexibility of program design be sacrificed to conform to common outcomes?

- What if common measures show no results? Then what?

- Who will be making the inferences about the results?

- How will results be communicated to stakeholders and clientele?

- Will having a common measure result in "teaching to the test"?

References

Adler-Baeder, F., Kerpelman, J., Griffin, M. M., & Schramm, D. G. (2010). Evaluating multiple prevention programs: Methods, results, and lessons learned. *Journal of Extension* [On-line], 48(6) Article 6FEA1. Available at: http://www.joe.org/joe/2010december/a1.php

AZ REACH (2014). CYFERNET Search Psychometrics. Retrieved from: https://cyfernetsearch.org/sites/default/files/PsychometricsFiles/Parent%20Education%20Survey%20(parents%20of%20children%20all%20ages)%20(cornell%20extension)_0.pdf

Baker, P., & Mott, F (1989). *NLSY child handbook 1989*. Center for Human Resources Research. Ohio State University, Columbus.

Boyd, H. H., (2009). Practical tips for evaluators and administrators to work together in building evaluation capacity. *Journal of Extension* [On-line], 47, (2). Article 2IAW1. Available at http://www.joe.org/joe/2009april/iw1.php

Cornell Cooperative Extension, (2009). *Parent education survey*. Retrieved from: https://cyfernetsearch.org/sites/default/files/PsychometricsFiles/Parent%20Education%20Survey%20

(parents%20of%20children%20all%20ages)%20(cornell%20extension)_0.pdf

Furco, A., Muller, P., & Ammons, M. S., (1998). *The civic responsibility survey*. Developed at the Service-Learning Research & Development Center, University of California, and Berkeley.

Mancini, J. A., & Marek, L.I . (2004). Sustaining community-based programs for families: Conceptualization and measurement. *Family Relations*, 53, 4, 339-347.

Payne, P. B. & McDonald, D. (2012). Using common evaluation instruments across multi-state community programs: A pilot study. *Journal of Extension* [On-line], 50,(4). Article RIB4. Available at: http://www.joe.org/joe/2012august/rb2.php

Payne, P. B., & McDonald, D. A. (2014). The importance of understanding dosage when evaluating parenting programs: Lessons from a pilot study. *Journal of Human Sciences and Extension*, 2(2). Retrieved from: http://media.wix.com/ugd/c8fe6e_2024a423ed064eaa8fb288aeb7ed1038.pdf

Roucan-Kane, M. (2008). Key facts and key resources for program evaluation. *Journal of Extension* [On-line], 46(1) Article 1TOT2. Available at: http://www.joe.org/joe/2008february/tt2.php

Santos, J.R.A. (1999). Cronbach's Alpha: A tool for assessing reliability of scales. *Journal of Extension*

[On-line], 37(2) Article 2TOT3. Available at: http://www.joe.org/joe/1999april/tt3.php

Spoth, R., Redmond, C., Haggerty, K., & Ward, T., (1995). A controlled parenting skills outcome study examining individual differences and attendance effects. *Journal of Marriage and Family,* 57, 2, p. 449-464.