Relationships between Quantitative Measures of Evaluation Plan and Program Modeling Quality and a Qualitative Measure of Participant Perceptions of an Evaluation Capacity Building Approach

Jennifer Brown Urban and Marissa Burgermaster

Montclair State University

Thomas Archibald

Cornell University

Alyssa Byrne

Montclair State University

Author Note

Jennifer Brown Urban, Family and Child Studies, Developmental Systems Science and Evaluation Research Lab, Montclair State University; Marissa Burgermaster, Developmental Systems Science and Evaluation Research Lab, Montclair State University; Thomas Archibald, Cornell Office for Research on Evaluation, Cornell University; Alyssa Byrne, Developmental Systems Science and Evaluation Research Lab, Montclair State University. This research was supported by NSF grant number 0814364.

Marissa Burgermaster is now at Department of Health and Behavior Studies, Teachers College Columbia University. Thomas Archibald is now at Department of Agricultural and Extension Education, Virginia Tech.

Correspondence concerning this manuscript should be sent to Jennifer Brown Urban, Montclair State University, Department of Family and Child Studies, 1 Normal Avenue, UN 4144, FCST, Montclair, NJ 07043. Email: urbanj@mail.montclair.edu.

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Abstract

Despite a heightened emphasis on building evaluation capacity and evaluation quality, there is a lack of tools available to identify high quality evaluation. In the context of testing the *Systems Evaluation Protocol (SEP)*, quantitative rubrics were designed and tested to assess the quality of evaluation plans and models. Interview data were also collected and analyzed using *a priori* codes. A mixed methods approach was used to synthesize quantitative and qualitative data and explore trends. Consistencies between data types were found for attitude and capacity, and disconnects were found for knowledge, cyberinfrastructure, time, and quality. This approach to data integration represents a novel way to tap the generative potential of divergence that arises when different methods produce contradictory results.

KEYWORDS: Evaluation Capacity Building, Evaluation Plan Quality, Logic Model Quality, Mixed Methods, Systems Evaluation

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Relationships between Quantitative Measures of Evaluation Plan and Program Modeling Quality and a Qualitative Measure of Participant Perceptions of an Evaluation Capacity Building Approach

Questions about evaluation quality have long pervaded the field and profession of evaluation. In her 2010 American Evaluation Association Presidential Address, Cooksy both reviewed and advanced the discussions around such questions, focusing especially on issues related to evaluator competency, the evaluation environment, and the supportive resources available in the evaluation community (Cooksy & Mark, 2012). Both she and Mark—who offers commentary on Cooksy's address—agree that knowing quality evaluation when you see it is not an easy task. There are some resources, such as the checklists available through the Western Michigan University Evaluation Center website (Stufflebeam, 1999, *inter alia*), which can guide one towards and help one assess evaluation quality, yet the need for more such resources persists. Similarly, the burgeoning subfield of evaluation capacity building (ECB) has increasingly focused on the notion of evaluation quality, since the inherent purpose of ECB is to help people (program implementers, usually) do higher quality evaluation. While much has been written about how to promote evaluation capacity, very little guidance has been offered on how to assess it (Labin, Duffy, Meyers, Wandersman, & Lesesne, 2012; Preskill & Boyle, 2008). In brief, despite the heightened emphasis on evaluation quality over the years, there is still a lack of tools available to identify high quality evaluation.

The need to assess the quality of evaluation is salient to many stakeholders, representing multiple levels of an organizational system. On one level, professional external evaluators could be expected to routinely conduct reflexive metaevaluations during and after any evaluation in which they are involved. In ECB contexts, assessment of evaluation quality provides evidence of

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

the efficacy of ECB efforts; thus, the implementers of such ECB efforts could be expected to collect data on the quality of the internal evaluations performed by program staff involved in their ECB initiatives. From a systems perspective, representatives of multiple hierarchical levels of the system in which an evaluand is nested could also be expected to have an interest in the quality of any evaluation conducted within that system. For example, a system such as Cooperative Extension within a given state has a central administration yet is also characterized by significant decentralization when it comes to evaluation; as such, the central administration may be interested in assessing the quality of evaluation being planned and implemented by its associated programs across the state. Similarly, a funding agency such as the United Way or the National Science Foundation could understandably be interested in knowing not just that their grantees are doing evaluation of funded programs, but also about the quality of that evaluation. Funding agencies that receive many competing applications to a request for funding (RFP) would also likely be interested in a formalized assessment of the quality of the various submitted evaluation plans and models.

The work presented in this paper contributes to the body of knowledge about evaluation quality, especially as it pertains to ECB efforts. It also presents a set of tools designed to assess the quality of evaluation plans and logic models. The quality of logic models and evaluation plans is an under-explored “mid-term” indicator of evaluation quality (and by extension, of evaluation capacity). Often, an evaluation can take months or even years to conduct. If one must wait until an evaluation is concluded, important opportunities to improve the quality of that evaluation (for example, with targeted ECB interventions) will be missed. Using evaluation plans and program logic models as early indicators of evaluation quality or evaluation capacity thus represents a novel contribution with potentially wide applications within the field of evaluation.

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Specifically as they pertain to ECB, the tools presented below offer additional benefits: Among the few existing tools designed to assess evaluation capacity, most focus on structural aspects of an organization and rely on self-reported attitudes of staff (Botcheva, White, & Huffman, 2002; Suarez-Balcazar et al., 2010; Volkov & King, 2007). Given this, we saw the need to develop tools that could more objectively provide evidence of the actual quality of evaluation plans and logic models. As we present in much more detail below, we recognized that a mixed methods approach would be essential in our efforts to develop and test a set of tools that could meet the needs outlined above.

Quantitative measures that provide early indicators of evaluation quality do already exist. For example, the Program Accountability Quality Scale (PAQS; Poole, Nelson, Carnahan, Chepenik, & Tubiak, 2000) is a quantitative measure of the quality of proposed performance measurement systems. For the current study, the goal was to develop a quantitative measure of the quality of evaluation plans, logic models, and pathway models and analyze the results of these quantitative measures in conjunction with qualitative data obtained from semi-structured interviews with program practitioners who participated in a specific evaluation capacity building approach known as the Systems Evaluation Protocol (Trochim et al., 2012). By including both quantitative and qualitative measures, we aim to gain a richer understanding of the relationship between a relatively more objective measure of quality and participants' reflections regarding several factors including their: attitudes toward evaluation, evaluation capacity, evaluation knowledge, sense of quality, and time needed to conduct evaluation planning activities. The quantitative measures developed for this study provide a unique approach to more objectively measure evaluation plan and model quality, while the qualitative measures provide unique insight into the factors that may enhance or hinder evaluation capacity building efforts.

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

This paper describes: (1) The development of a set of rubrics designed to assess the quality of evaluation plans, logic models, and pathway models, (2) The quantitative testing of those rubrics' reliability and internal consistency, (3) The development of a qualitative interview data collection and analysis protocol, and (4) The mixed methods synthesis of the quantitative data on evaluation plan, logic model and pathway model quality with qualitative interview data to explore trends across the qualitative and quantitative data particularly as they relate to evaluation capacity building efforts. We propose that our approach to mixed methods design and analysis—specifically the use of a linear regression analysis representing quantitative data to guide and strengthen our analysis of qualitative interview data—presents a noteworthy contribution to the field of mixed methods research and evaluation. We see this approach to data integration as a novel addition to the mixed methods tradition of generatively yielding new insights and understandings which would otherwise be missed if only quantitative or qualitative data and analysis are used (Greene, 2008).

Context

The Evaluation Plan (EP), Logic Model (LM) and Pathway Model (PM) Rubrics described here were developed as part of the Systems Evaluation Protocol (SEP), a step-by-step guide for program practitioners and evaluation professionals who wish to integrate a systems thinking perspective into their work (Trochim et al., 2012). The SEP is a systems-based approach, designed to build the internal evaluation capacity, including evaluative thinking, of program practitioners and administrators. Specifically created in the context of education and outreach programs in Science, Technology, Engineering, and Mathematics, the SEP is intended to be generally applicable to any type of program evaluation. The SEP is currently being tested as part of a five-year longitudinal cohort sequential study, in which participating programs take

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

part in a facilitated version of the SEP in partnership with the Cornell Office for Research on Evaluation (CORE).

During the Evaluation Planning phase of their Evaluation Partnership with CORE, program practitioners ("partners") develop an evaluation plan, including program-based logic and pathway models. This plan is intended to serve as a guide for the implementation of the evaluation and includes broad evaluation questions, specific sampling strategies, the identification or development of measures to assess the evaluation questions, an evaluation design, an analysis plan, a reporting plan, and an implementation plan and schedule. Program practitioners receive feedback from CORE in the form of the EP, LM, and PM Rubrics and are encouraged to use this feedback to revise their plan before beginning evaluation implementation.

When evaluating the efficacy of the SEP, the EP, LM, and PM Rubrics were included as measures to assess whether there is any systematic relationship between utilization of the SEP and indicators of quality evaluation planning (i.e., high quality EPs, LMs, and PMs). Revisions were made to the rubrics in order to reflect the change in purpose from a feedback tool to a summative scoring tool allowing quantification of the quality of participating programs' EPs, LMs, and PMs. This article presents these rubrics including preliminary assessments of reliability and validity. In addition, this article describes the results of a mixed methods analysis using the quantitative rubric measures and qualitative interview data to assess evaluation capacity building using the SEP, highlighting the ways in which our data integration approach contributes to ongoing discussions in the mixed methods literature (Fielding, 2012; Mertens & Hesse-Biber, 2012).

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Quantitative Methods (Rubrics)

The EP, LM, and PM Rubrics presented here were initially designed as feedback tools to help participating programs improve their evaluation plans, logic and pathway models before beginning evaluation implementation. The organic development of subsections and scale items is based upon the SEP and are reflective of the rubrics' origins as feedback measures. Revisions in language and scale were made to reflect the change in the rubrics' purpose for use as a quantifiable measure applicable across evaluation contexts. When using the rubric as a feedback tool, the numerical scores are not necessarily shared with partners. Rather, a focus on qualitative feedback in the form of comments in each section is maintained; numerical scoring is primarily completed for comparative purposes.

Rubric Testing during Development

The EP, LM, and PM rubrics were assessed to determine if rubric item phrasing and scale allowed consistency across rater pairs. In two rounds of testing using two randomly selected programs, six research team raters rated both programs using the EP, LM, and PM Rubrics. Reliability testing, including Cohen's Kappa, Pearson product moment correlations, intra-class correlation, and percent agreement were conducted. The initial testing resulted in the elimination of items, the rephrasing of items, and the separation of items. The wording of the rating scale was revised for precision. During this testing, subsection scores were determined to be more reliable than individual item scores.

Next, tests of rubric inter-rater reliability were conducted to determine overall rubric reliability for individual items and sub-section scores using 12 randomly selected programs. Non-repeating rater pairs were randomly assigned to rate the programs. Reliability testing, including Pearson product moment correlations, intra-class correlations, and weighted and un-

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

weighted Cohen's Kappa was conducted. Results yielded additional revisions to the rubrics, including another scale wording revision, a change to a 0-4 scale, and the revision of item verb tenses. Final tests of inter-rater reliability and internal consistency were conducted on the finalized rubrics.

Rubric Structure

The EP, LM, and PM rubrics were designed so that they could be used as separate measures. Therefore, the structure of each rubric is described below. For this specific study, scores on the LM and PM Rubrics were summed to create a composite LM/PM score.

Evaluation Plan Rubric. The EP Rubric (Appendix A) is a 48-item measure, divided into 12 sub-sections as follows: Program Mission/Purpose Statement (n=2), Program Description (n=4), Evaluation Purpose (n=4), Evaluation Questions (n=4), Sampling (n=6), Measurement (n=7), Design (n=4), Data Collection and Management (n=3), Data Analysis (n=2), Evaluation Reporting and Utilization (n=4), Evaluation Timeline (n=3), Overall (n=4). Each sub-section contains specific items related to the sub-section topic and are rated from "Unacceptable" to "Excellent" on a 0-4 Likert-type scale. The rubric is designed to follow the typical structure of an evaluation plan and encourages the rater to consider appropriate evaluation design for the program setting and stage of development. For example, evaluation questions should be appropriate given the stage of development of the program and its prior evaluations (i.e. design, measurement, analysis, and reporting should all be aligned to the evaluation questions). The overall section, located at the end of the EP Rubric, emphasizes the importance of an EP as a communication tool and addresses evaluation capacity at its most basic level: not only must sample, design, measurement, analysis and reporting plans be aligned with evaluation questions and with each other, the evaluation plan must be feasible.

Logic Model Rubric. The LM Rubric (Appendix B) is an 18-item measure**,** divided into 7 sub-sections as follows: Inputs (n=2), Activities (n=3), Outputs (n=3), Outcomes (n=4), Assumptions (n=2), Context (n=2), and Overall (n=2). Like the EP Rubric, each item within each sub-section is rated from "Unacceptable" to "Excellent" on a 0-4 Likert-type scale and each sub-section contains specific items related to the sub-section topic. Again, the rubric and sub-sections are organized to follow the typical structure of a logic model. The items emphasize consistency with the corresponding evaluation plan, and the phrasing, appropriateness, and comprehensiveness of the outputs and outcomes considering program context and evaluation scope. The Overall section, located at the end of the LM Rubric, gives the rater an opportunity to assess the clarity of the model and its effectiveness in communicating a summary of the program and its goals.

Pathway Model Rubric. A pathway model is similar to a logic model in that it is a visual depiction of a program's logic. However, it adds a significant element by incorporating the logical connections between specific activities and outcomes, thereby better telling the story of how the program works. Figure 1 provides an example of a pathway model. The PM Rubric (Appendix B) is a 10-item measure**,** divided into 3 sub-sections as follows: Items (n=1), Connections and Pathways (n=7), Overall (n=2). Again each item is rated from "Unacceptable" to "Excellent" on a 0-4 Likert-type scale and each sub-section contains specific items related to the sub-section topic. The items emphasize consistency with the corresponding logic model, and the sequencing, completeness, and logic of the connections among activities and various levels of outcomes. Similar to the EP and LM Rubrics, the Overall section encourages raters to holistically assess the PM as a communication tool.

<< Insert Figure 1 here >>

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Rubric Scoring

Rubric scores are derived by summing the rater's score for each sub-section. Analyses were conducted based on these summed sub-section scores and not based upon individual item scores. Initial testing confirmed a hypothesis that a higher level of agreement on the overall sub-section score existed despite greater variability at the individual item level.

Methods & Analysis

Statistical Analysis of Inter-Rater Reliability. Data for the rubric analyses comes from completed EP, LM, and PM Rubrics for programs participating in the study. Raters were randomly assigned to rate the plans and models for their assigned programs using the Rubrics. Initially, the inter-rater reliability of rubric scoring was assessed followed by tests of internal consistency. Fifteen randomly selected programs were used for inter-rater reliability testing and non-repeating rater-pairs were randomly assigned to programs. Inter-rater reliability of rubric sub-section scores was tested using Pearson product-moment correlations and intra-class correlations (ICC; Shrout & Fleiss, 1979). The inter-rater reliability of rubric sub-section scores for the EP Rubric ranged from ICC (2, 1) = .557 ($p < .05$) to ICC (2, 1) = .946 ($p < .001$). The average ICC across the 15 rater-pairs was .760 and all ICC values reached significance at $p < .05$. The inter-rater reliability of rubric sub-section scores for the LM Rubric ranged from ICC (2, 1) = .168 (not significant) to 1.00 ($p < .001$). The average ICC across the 15 rater pairs was .743 and most ICC values reached significance at $p < .05$ with the exception of two rater pairs whose ICC value was significant at $p < .10$ (ICC (2, 1) = .605 and ICC (2, 1) = .510) and one where the ICC value did not reach significance (ICC (2, 1) = .168). The inter-rater reliability of rubric sub-section scores for the PM Rubric ranged from ICC (2, 1) = .444 (not significant) to ICC (2, 1) = .999 ($p < .001$). The average ICC across 15 rater pairs was .935 and all ICC values reached

significance at $p < .05$ with the exception of one rater pair whose ICC value did not reach significance. ICC values greater than .40 are considered fair (Cicchetti, 1994). Across tests of all three rubrics, only two rater pairs had reliability scores that fell below .40. Based on these results, the research team determined that an acceptable level of inter-rater reliability had been established.

Internal consistency. After the establishment of acceptable rubric inter-rater reliability, an additional 24 programs were rated by a single rater who was randomly assigned to rate the program bringing the total number of programs used for assessment of internal consistency to 39. Cronbach's Alpha, inter-item correlations, and corrected item-total correlations were used to assess internal consistency.

The overall reliability estimate for the EP Rubric was $\alpha = .949$. Cronbach's alpha for each of the sub-scales was acceptable: program/mission statement ($\alpha = .932$), evaluation purpose ($\alpha = .745$), program description ($\alpha = .774$), evaluation questions ($\alpha = .701$), sampling ($\alpha = .826$), measurement ($\alpha = .831$), design ($\alpha = .805$), data collection and management ($\alpha = .888$), data analysis ($\alpha = .785$), evaluation reporting and utilization ($\alpha = .848$), evaluation timeline ($\alpha = .855$), and overall ($\alpha = .864$)

The overall reliability estimate for the LM Rubric was $\alpha = .923$. Cronbach's alpha for each of the sub-scales was acceptable: assumptions ($\alpha = .941$), context ($\alpha = .767$), inputs ($\alpha = .761$), outputs ($\alpha = .845$), activities ($\alpha = .611$), outcomes ($\alpha = .686$), and overall ($\alpha = .882$).

The overall reliability estimate for the PM Rubric was $\alpha = .822$. The items sub-section only has one item, so internal consistency for this sub-scale was not assessed. Cronbach's alpha for the connections and pathways sub-section was acceptable ($\alpha = .823$) while the overall sub-

section was low ($\alpha = .404$) indicating that it makes the most sense to treat the PM Rubric as a single scale without separate sub-sections.

Qualitative Methods (Interviews)

In addition to conducting quantitative assessments of evaluation plans, logic models, and pathway models, thirty minute phone interviews were also conducted with program practitioners and administrators at the completion of the evaluation planning phase of the SEP. Interviews with participants were conducted from an interpretivist paradigm.

Methods & Analysis

Interviews were conducted by trained interviewers in the Developmental Systems Science and Evaluation Research Lab at Montclair State University and the Survey Research Institute at Cornell University who had not previously interacted with participants in any way. The SEP facilitators from CORE were deliberately not engaged in the interview process in order to ensure the trustworthiness and credibility of the interviewers (Mertens, 2005). Interview questions included items about general impressions of evaluation, such as, “Based on your experience with evaluation planning in the Evaluation Partnership, what do you think promotes/hinders evaluation in an organization or program?” as well as items about the impact of the Evaluation Partnership on an organizational level, such as, “Can you describe any spillover effects your program’s involvement with the Evaluation Partnership may have had on other programs within your organization?” Interviews were conducted by trained research assistants. Balancing feasibility and precision, interviews were not audio-recorded, but extensive notes and verbatim quotes were typed during the interviews (Kvale & Brinkmann, 2009; Tessier, 2012). In order to improve accuracy of interview notes, interviewers conducted member checks with

participants by paraphrasing their responses during the interviews and reviewed interview notes immediately after each interview.

Participants. Interviews were conducted with personnel who held various positions within the program or organization and were engaged on some level with evaluation planning, including program practitioners and program administrators, depending on the structure of the program and organization. A total of 60 interviews were used for the current analysis. Interviews were conducted with personnel from 31 of the 39 programs included in the quantitative rubric analysis and 19 of these programs had interview data from multiple interviewees.

Coding. Research assistants who received extensive training in qualitative data analysis (QDA) and *NVivo10* (QSR International, 2012) coded the interview data using nodes corresponding to a coding dictionary.

***Coding dictionary*.** The research team worked together to adapt the project variables and research questions into *a priori* codes. The relevant research questions focused on program delivery and outcomes for program participants. These codes were defined and examples were provided in a coding dictionary designed for this project. The *a priori* codes used as part of the current analysis are: Netway (the application and use of the Evaluation Partnership's cyberinfrastructure), Time (time in reference to the Evaluation Partnership), Knowledge (knowledge, knowledge gained, or lack of knowledge before participating in the Evaluation Partnership), Attitude (the opinion, feeling, or attitude of a participant), Capacity (resources available for evaluation within a program or organization), and Quality (better or worse product as a result of participating in our program).

Coding Process. Completed interview notes were randomly assigned to two research assistants who were instructed to read each interview multiple times and assign chunks of text to

the *a priori* code nodes as appropriate. In addition, nodes were created for all participant identification numbers as well as demographic information (e.g. participant job title) to facilitate the mixed methods analysis.

Inter-coder reliability. Four rounds of inter-coder comparison queries were conducted until good inter-coder reliability was achieved between the two research assistants. During each round the research assistants would each code five interviews using the *a priori* coding dictionary. Then, an inter-coder comparison query (including percent agreement and Cohen's Kappa) was run in *NVivo* for the total number of interviews completed (i.e. round 1 n=5, round 2 n=10, round 3 n=15, round 4 n=20). After the first round, the dictionary was revised significantly. After each subsequent round, the definitions and coding examples were discussed and revised with the research team, including the coders. After four rounds, at n=20 interviews, Kappa scores for all codes ranged from 0.70-0.93, and the research team determined that adequate inter-coder reliability had been reached. Once inter-coder reliability was established, the remaining interview notes were randomly assigned to one of the two coders.

Mixed Methods Analysis

The data analysis followed a modified sequential explanatory mixed methods design (Hesse-Biber, 2010), wherein the quantitative data were collected and preliminarily analyzed separately. Then the quantitative rubric quality data were used as a framework within which to further analyze and interpret the qualitative interview data. Figure 2 depicts the mixed methods design employed in the current study.

<< Insert Figure 2 here >>

A linear regression analysis was conducted on the quantitative rubric data to test whether a sum score for logic and pathway model rubric scores predicted evaluation plan rubric scores. A

significant linear relationship was found, $F(1,37) = 29.57$, $p < .001$. Figure 3 presents a visual depiction of this linear relationship. Each program was given a unique identifier (e.g., P002) and programs are represented as points in the figure. Each point represents the summed LM/PM rubric score and the EP rubric score for a specific program. In general, as programs' LM/PM rubric scores improved, so too did EP rubric scores. The visual depiction of this linear relationship grounded the qualitative analysis.

<< Insert Figure 3 here >>

The mixed methods analysis itself was designed to be an iterative process, which included multiple rounds of discussion among members of the research team. The data for each of the *a priori* codes were analyzed within the framework of the quantitative rubric scores. Specifically, the research team sought patterns or trends in the data for each code among programs with similarly high or low rubric scores. The mixed methods analysis was conducted in several rounds for each of the a priori codes. An NVivo query was run on the interview data for each code and the data was organized by program according to the order established by the EP and LM/PM rubric regression line.

For ease of explanation, the example of the *a priori* code "attitude" will be used in the remainder of this description. First, all of the interview data that had been coded for "attitude" were retrieved and organized. The research team met and each team member received a single program's results for the code "attitude", beginning with the programs with the lowest quality EP and LM/PM scores. In other words, the research team began by analyzing the qualitative data that corresponded with the point (program) at the lower left hand corner of the regression line presented in Figure 3. Each team member read through the parts of their assigned program's interview transcript coded "attitude" and summarized the general themes that emerged. Team

members then discussed their summaries, noting common themes or striking differences and recording one or more quotes from the transcripts that seemed reflective of that set of transcripts. Then, the programs with the next highest EP and LM/PM rubric scores became the focus of analysis. The qualitative query results for "attitude" from this group of programs was distributed among team members who read and summarized the parts of their assigned program's interview transcript coded "attitude." The summaries were discussed among the team members and common themes and differences were again noted, along with exemplar quotes. This process continued until all of the programs with interview transcripts coded for "attitude" were summarized and discussed. At this point, the team members discussed notes from all rounds of the mixed methods analysis for "attitude". Trends across the rounds of analysis were discussed. In particular, team members sought to determine if there were notable differences or similarities between lower and higher scoring programs. This discussion was then summarized, along with the notes and exemplar quotes from the earlier rounds of the analysis, and several team members reviewed the write-up to ensure that it was reflective of all the programs included in the analysis. In some cases, the code summary was broken down into code subtopics, (e.g. the "Attitude" summary was separated into "Optimism about Evaluation" and "Valuing Evaluation"). As a concluding step, the team reviewed summaries for accuracy and consensus. This process was repeated across all of the programs with available interview data for each of the *a priori* codes: attitude (n = 23), knowledge (n = 22), Netway (n = 30), quality (n = 13), capacity (n=28), and time (n = 31).

Results

Trends in the qualitative data corresponding to increasing rubric quality scores as well as the absence of such trends were noted. The results of this analysis are presented below for each of the *a priori* codes.

Consistencies in Qualitative and Quantitative Trends

Consistency refers to an agreement or accordance between the qualitative and quantitative data. In other words, these are cases where the qualitative data reflected the pattern of results found for the quantitative rubric quality scores. For example, the qualitative data for programs with higher rubric scores (indicating better quality) reflected more positive experiences while the qualitative data for programs with lower rubric scores (indicating lower quality) reflected more negative experiences. Consistencies are discussed in terms of the trend of rubric quality scores, from low to high as depicted in Figure 3.

Attitude. For the purposes of this research, "attitude" refers to the opinion, feeling, or attitude of a participant. In general, the interview excerpts coded for attitude became more positive as a program's EP, LM and PM quality improved. This trend is described below in terms of two sub-codes for attitude: (1) optimism about evaluation, and (2) valuing evaluation.

Optimism about evaluation. A common trend throughout the qualitative data analysis was that higher scoring programs tended to have a more positive and optimistic attitude about evaluation compared to lower scoring programs. Many staff from lower scoring programs thought the evaluation planning process was difficult, daunting, and frustrating. A program practitioner from a lower scoring program (P005) stated, "The process is so damn long - it takes a lot of time. If we were going to evaluate everything we did, it would take years and years and years. It's hard to find the time to just evaluate one project. We're not so interested in the

research and the theory aspect of it." In comparison, staff from higher scoring programs frequently acknowledged the arduousness of evaluation planning, but many found the overall process to be positive. A program practitioner from a higher scoring program (P018) stated, "I was really surprised how much I enjoyed [the process]. I would have thought evaluation was right up there with going to the dentist."

Valuing evaluation. Program practitioners from lower scoring programs seemed to have a difficult time convincing other staff of the importance of evaluation. Many program practitioners from lower scoring programs indicated it was difficult getting staff members to view and understand the significance of evaluation. A program administrator from a lower scoring program (P015) stated, "It was hard to convince our newer staff that this was valuable." Having staff that did not value evaluation and saw it as insignificant made it difficult for program practitioners to incorporate evaluation into their work, frequently resulting in frustration. A program practitioner from a lower scoring program (P003) indicated "a lack of perceived need or importance – it is too easy to get caught up in our day-to-day obligations and lose sight of planning – and a lack of organizational commitment." In comparison, higher scoring programs placed value on evaluation, and promoted it as important. As stated by a program administrator from a higher scoring program (P020), "I'm used to thinking about evaluation as something you do last, but putting a priority on it helps a lot as an organization."

Capacity. For the purposes of this research, "capacity" refers to resources available for evaluation within a program and/or organization. This includes human resources, such as experience, external resource and leadership support, as well as more concrete resources, such as tools and funding. In general, as a program's EP, LM and PM rubric quality scores improved, the program participants' reports of the program's capacity for evaluation tended to improve as

well. Capacity in terms of human resources (e.g., personnel evaluation experience and efficacy, the use of external human resources) follows the quantitative trend. Although many programs of differing quality remarked on the importance of leadership and colleague support in terms of capacity, the higher scoring programs were more likely to imply the actual presence of leadership and colleague support. Similar trends were less explicit for the more concrete resources, like tools and funding, which were mentioned more frequently by lower scoring programs, and for the capacity for funding and stakeholder communication, which was mentioned more frequently by higher scoring programs.

Evaluation experience and efficacy. A common theme among program practitioners was the value of human resources, including personnel with evaluation experience and confidence in conducting evaluations. Across low and high scoring programs it was acknowledged that having personnel who have experience with evaluation improves that program's evaluation capacity: "A lot of it is having a basic understanding of tools for evaluation and plans" (P008).

Many lower scoring programs indicated the importance of having personnel with evaluation experience but may have lacked staff that had this evaluation expertise. Program practitioners from lower scoring programs indicated the need for personnel with evaluation experience: "[it would be beneficial to have] somebody in the organization that has some level of evaluation experience to already know that it is important and believe in the importance of it" (P040). However, they did not necessarily indicate capacity in that area.

It is possible that staff from higher scoring programs had more prior experience with program evaluation in comparison to staff from lower scoring programs. Staff from higher scoring programs tended to report having evaluation experience or a colleague with evaluation

experience more than staff from lower scoring programs. Staff from higher scoring programs referenced both prior evaluation knowledge, "The one person who really understood things…had taken grad level work in evaluation" (P010), as well as evaluation experience gained during participation in the SEP, "Our discussions at staff meetings have been really productive since our trained staff member now is bringing her knowledge to meetings and using what she's learned, her perspective is really helping us all" (P020). These program participants also referenced their own increased capacity. For example, a program practitioner from a higher scoring program (P019) stated, "I know that I have a much better sense of how to do evaluation – the whole step by step process. I think once we put it in place we'll learn a lot about how to improve and otherwise modify our program."

Experience with evaluation, whether it was prior experience or experience during participation in the SEP, may be related to evaluation efficacy. Lower scoring programs generally indicated lower evaluation efficacy than higher scoring programs. For example, a program practitioner from a lower scoring program (P021), referenced his "lack of confidence in evaluation and…skills to do evaluation," whereas a program practitioner from a higher scoring program (P027) reported that he "came out of [the SEP] with a lot more knowledge and confidence and more concrete tools."

Support of evaluation. Another frequently referenced human resource was the support of evaluation experts, such as Evaluation Partnership facilitation staff at CORE, as well as the support of colleagues and program or organization leadership.

Many program practitioners commented on how support from evaluation professionals at CORE helped to increase their capacity; however, there was a prominent difference in higher and lower scoring programs in terms of practitioners' use of CORE staff support during evaluation

planning. As a program practitioner from a lower scoring program (P040) stated, “I didn’t really take advantage of picking up the phone and calling the staff at CORE. And I think maybe if, I don’t know, it would have helped if that were required instead of optional, to call the staff at CORE. Having the personal connection was really helpful.” Higher scoring programs seemed to place greater emphasis on support from the CORE staff and reported using CORE support more than lower scoring programs. The on-going support seemed to help programs decide what measures and tools to use in their evaluation. A program practitioner from a higher scoring program (P018) stated, “I think the most beneficial part has been being able to access the expertise of people who know how to do this, they even have worked with us one on one to refine our plans and think about what to look for, think better about how we’re doing this project. I don’t know that we could do it without them – staff at CORE.” By utilizing professional expertise, programs may have increased both their capacity and the quality of their evaluation plans.

Both low and high scoring programs indicated the importance of supportive program and organization leadership and colleagues. A lower scoring program (P021) indicated, “Executive director support would be nice.” Some higher scoring programs explicitly reported leadership support, “Having our director’s support certainly helped a lot too” (P018). In some cases, the leadership’s support of evaluation was made known in more indirect ways, such as including evaluation duties in specific job descriptions (P014).

For several lower scoring programs, a lack of support from colleagues and staff turnover seemed to affect evaluation capacity. Staff turnover, especially, produced frustration and led programs to focus less on evaluation. A program administrator provides an example: “…we had two changes in staffing. So the person who was originally in charge retired last year, and then we

had an interim director who also left. In the meantime, we also had a grant that was due, and we had asked for an extension on the application because of losing our director and the extension was denied. So the whole evaluation planning process became unusable with the staff changes and with the fact that we were so pushed for time on the grant that we couldn't put any time into evaluation" (P013).

Concrete resources. Comments regarding limitations in funding and other concrete resources were nearly universal across the EP, LM, PM quality scale, "We're asked to do more with less staff and less funding and less of everything every day" (P005). Only one program, P018, one of the highest scoring programs on the quantitative quality measures indicated that the program had adequate resources and tools for evaluation capacity.

Funding and stakeholder communication. Similar to findings regarding concrete resources, the benefits of participation in the SEP for securing funding and stakeholder communication was common across all levels of EP, LM, and PM quality. For example, a program practitioner reported that his program would have a very sound evaluation document that could be given to supporters, funders, and administration (P015) and a program administrator pointed out that, "We also have multiple funding partners and these…often have no understanding of what we do. So we have to demonstrate our value to them in terms of showing them outcomes and framing those outcomes in terms that they understand" (P027).

Disconnects between Qualitative and Quantitative Trends

Disconnects refer to an inconsistency between the qualitative and quantitative data. These are cases where the qualitative data did not reflect the pattern of results found for the quantitative rubric quality scores. Disconnects are discussed in terms of the trend of rubric quality scores, from low to high as depicted in Figure 3. As we reiterate in the discussion section below, one of

the primary benefits of the mixed methods data integration approach presented in this paper is that it moves beyond triangulation—which often focuses solely on convergence and confirmation of results from multiple methods (Greene, 2007)—to actively pursue deeper understanding through its emphasis on divergence and dissonance. Triangulation is a topic that has deservedly received increased scholarly attention within the mixed methods field in recent years. The contributors to a recent special issue of this journal "raise many questions about the meaning of triangulation, its philosophical positioning in the mixed methods community, and strategies for using triangulation in the design of mixed methods studies, analysis and interpretation of data, and making visible subjugated voices" (Mertens & Hesse-Biber, 2012, p. 75). This article adds to that discussion by presenting an empirical example of a technologically-assisted data integration approach (Fielding, 2012) that was designed with (conventional definitions and purposes of) triangulation in mind, yet which led us, as our analysis unfolded, to a more complex portrait of our phenomena of interest, "including the peaks and valleys of both dissonance and consonance" (Greene, 2007, p. 103).

Knowledge. For the purposes of this research, "knowledge" refers to reports of existing knowledge of evaluation, knowledge of evaluation gained through participation in the SEP, or lack of knowledge of evaluation. Although there were no distinct trends in knowledge corresponding to the quantitative framework, most program practitioners indicated an increase in knowledge after participation in the SEP: "My skills and knowledge of how to prepare for evaluation have improved tremendously" (P003). Many program staff acknowledged that participation in the SEP was associated with better understanding of how to prepare an evaluation plan, increased skills and knowledge about evaluation, and increased understanding of the use of logic models. Some staff noted that learning new theoretical knowledge and

vocabulary was difficult and cumbersome but they had a better understanding of it as they progressed through the evaluation process.

Furthermore, in addition to gaining knowledge, program practitioners changed their thinking about evaluation from being a one-step process to a more fluid and iterative process. A program administrator stated, "For my staff this training took their understanding of evaluation to a much deeper level and helped them understand how to build evaluation into the program design and into the grant applications and contracts for programs" (P002).

Netway (cyberinfrastructure). For the purposes of this research, "Netway" refers to references to the cyberinfrastructure that was used in conjunction with the SEP. Although there were no distinct trends in the qualitative data corresponding to the quantitative framework, many program practitioners indicated that the Netway was helpful in that it allowed them to see other programs' ideas and models as well as potential partners and collaborators. The Netway is designed so that when evaluators enter their program's information, the system will identify other existing programs that have common elements and themes. A program practitioner stated, "I used the Netway a lot…It was the first time I'd ever done Pathway models. It was really interesting to me and I found the most useful was that I could look at other groups' work and see what they had done. I think for our program the Netway was the most useful as an organizational tool for our thought processes" (P016). Many program practitioners and administrators reported the benefits of seeing activities, outputs, and outcomes of other programs. Viewing these shared outcomes enabled many practitioners who were developing evaluation plans and models to adopt certain elements for their own projects.

Program practitioners also reported the Netway was beneficial for creating and visualizing logic models. A program practitioner stated, "I really liked the ease in which you can

use the logic model, that particular image really comes to life very easily. It was primarily the repository for getting the evaluation plan put together" (P040). Overall, the Netway seemed to help program practitioners become more specific and detailed in their thinking.

Time. For the purposes of this research, "time" refers to references to time (e.g., quantity of time or timing) in regard to how the SEP was delivered. Although there were no distinct trends corresponding to the quantitative framework, most program practitioners indicated that the evaluation planning process is time-consuming and arduous. One program practitioner indicated, "I felt really overwhelmed in the beginning simply because of how involved it was. I really didn't believe at first that this was going to work out for me because it involved so much time" (P022).

Quality. For the purposes of this research, "quality" refers to better or worse products as a result of participating in the SEP. There were a limited number of passages coded for quality and those coded did not indicate a trend corresponding to the quantitative framework. However, program practitioners reported benefits to the quality of their evaluation planning as a result of the Evaluation Partnership. Despite developing lower scoring EPs, LMs and PMs, staff from lower scoring programs reported improved quality in both evaluation and program planning through improved logic models, program adjustment, and improved communication with team members. A program practitioner stated, "The most obvious benefit to our program was really getting some hard data out of this whole process about the outcomes of our program and we really are using different measures that we had before. So I expect much better information about our impacts" (P016). Additionally, many staff across the range of EP, LM, and PM rubric quality scores stated that they learned how to better serve their target audience and they developed a

quality evaluation plan or annual report that could be given to funders, supporters, and administrators.

Discussion

In complex, dynamic, multifaceted endeavors such as evaluation planning and evaluation capacity building, assessing and understanding quality are not trivial tasks. The rubrics are tools designed to assess the quality of evaluation plans, logic models, and pathway models. We have shown that these tools are both valid and useful and we see the potential for the rubrics to be disseminated and implemented in other ECB and evaluation planning contexts. As the results presented above indicate, the mixed methods approach implemented in this research project provided a number of insights about evaluation quality and evaluation capacity. Our use of mixed methods served multiple purposes. On one level, the interaction between the qualitative interview data and the quantitative rubric data allowed us to better assess the validity and utility of the quantitative tools—this purpose of using qualitative data to refine the development of quantitative measures is quite common (Teddlie & Tashakkori, 2006). On another level, we posit that, because this project used mixed methods, we were able to uncover new patterns and structures of meaning pertaining to assessment of evaluation quality that we would have neglected if we only used the rubrics. This ability to generatively yield new insights and understandings is one often discussed advantage to using mixed methods (Greene, 2008).

At this level, our study is an example of a dialectic stance on mixing paradigms while mixing methods, in that the paradigmatic differences between regression analysis of quantitative data and interpretive analysis of qualitative data are "respectfully and intentionally used together to engage meaningfully with difference and, through the tensions created by juxtaposing different paradigms, to achieve dialectical discovery of enhanced, reframed, or new

understandings" (Greene, 2007, p. 69). Our data integration approach allows for the juxtaposition of different perspectives, where such difference is constitutive and generative (Greene, 2005). As such, while we initially had purposes of triangulation in mind, we ultimately arrived at a novel manifestation of what Greene (2007) calls the "initiation" purpose for mixing methods, in which "different methods are implemented to assess various facets of the *same complex phenomenon*" to identify divergence or dissonance (p. 103, emphasis in the original). Such mixing yields what Tom Cook refers to as an "empirical puzzle"—a puzzle that can simultaneously reveal new insights and point the way towards further inquiry that may be required (Greene, 2007).

For example, regarding the very notion of quality (and, importantly, how to gather good data about quality) we observed that all study participants believed (according to their interview responses) that they did higher quality evaluation planning work after participating in the SEP, despite the fact that the rubric scores indicated high quality work for some, but not all participants. On one hand, these apparently discrepant results suggest that tools such as the rubrics are especially helpful measures because they provide a more objective impression of evaluation quality than do self-reports. At the same time, the disconnect between objective and subjective interpretations opens up further lines of inquiry, for example, about how to understand and compare relative gains in capacity made within an organization, or about the ways self-perception of capacity relate to outwardly demonstrable indications of capacity. Due to a limitation of this study, however—the fact that it is a cross sectional comparison as opposed to a longitudinal assessment—these interpretations must be tempered. Because of this shortcoming in the study design, we were not able to ascertain changes in the quality of participants' logic models, pathway models, and evaluation plans over time. It is possible that all participants did in

fact make gains in quality, but because some were starting at relatively lower levels of evaluation capacity, those gains still did not result in work that was rated as high-quality using the rubrics.

Additional notable findings and new structures of meaning that emerged from the relationship between the quantitative rubric data and the qualitative data involve the notions of time, attitude, and capacity. All three of these factors are often discussed in the ECB literature. Hence, this study both corroborates and contributes to work in that field. Regarding time, almost all study participants reported in interviews that the process was very time consuming, regardless of whether they represented high- or low-scoring programs (as assessed using the rubrics). This suggests that, while time is clearly an important factor in ECB and in evaluation, time limitations should not preclude an organization from doing high quality evaluation. These findings point to a need to better understand, through further research, what other factors (such as attitudes, strategies, processes, etc.) are associated with higher scoring programs ability to overcome the time limitations facing most organizations. Our findings relative to attitude corroborate other ECB studies; as could be expected, higher scoring programs often had more positive attitudes. This finding reiterates the need for ECB initiatives to work intentionally and persistently to try to effect more positive attitudes about evaluation among the people with whom they work. Similarly, one aspect of capacity that emerged from the mixed methods analysis was that higher scoring programs were more likely to reach out to CORE staff and take advantage of one-on-one support and coaching, whereas lower scoring programs knew about that help and retrospectively regretted not seeking it out more regularly. This suggests something like a positive feedback loop, in which programs that have some capacity and have more positive attitudes about evaluation tend to more actively seek out advanced help on evaluation, resulting in higher quality work. Although these correlative interpretations cannot elaborate causal pathways which could

trace exactly how to best help lower scoring programs improve their evaluation capacity, they do provide novel insights about evaluation quality. Finally, as discussed below, all of these findings point to potentially fruitful areas of further study.

Although the rubrics provide an initial advance toward objectively measuring evaluation plan and model quality, additional work is needed to further validate these tools. For example, a larger and more diverse sample of completed rubrics is needed in order to conduct factor analyses. A more thorough assessment of the factor structure may also indicate a need to apply a weighting scheme to the various sub-sections. In addition, the programs that were included in the current analyses were all engaging in a project that incorporates elements of systems thinking into evaluation planning. The rubrics were intentionally designed to be used with programs using a variety of evaluation approaches (including systems evaluation and more traditional evaluation approaches). Future studies should explore whether variations in the approach to evaluation planning are associated with differences in evaluation plan and model quality.

Despite the need for additional research, the rubrics may already be a valuable resource particularly for those who are seeking early benchmarks of evaluation capacity. For example, funders may want to use the rubrics to compare the quality of plans and models submitted in response to a request for proposals. Those who teach evaluation planning may find the rubrics to be a useful way to provide both qualitative feedback and quantitative scores for students. Organizations may find the rubrics to be a useful tool for internally tracking evaluation planning quality. This study not only provides potentially useful quantitative measures of evaluation plan and model quality but it also presents a unique approach to mixed methods analysis that augments our understanding of ECB beyond what we could have learned from using either

quantitative or qualitative methods in isolation, thus providing an empirical example of the generative potential of dialectical paradigm mixing through a novel approach to integrated data analysis.

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

References

Botcheva, L., White, C. R., & Huffman, L. C. (2002). Learning culture and outcomes measurement practices in community agencies. *American Journal of Evaluation, 23*(4), 421-434.

Cichetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Cooksy, L. J., & Mark, M. M. (2012). Influences on evaluation quality. *American Journal of Evaluation, 33*(1), 79-87.

Fielding, N. G. (2012). Triangulation and mixed methods designs: Data integration with new research technologies. *Journal of Mixed Methods Research, 6*(2), 124-136.

Greene, J. C. (2005). The generative potential of mixed methods inquiry. *International Journal of Research & Method in Education, 28*(2), 207-211.

Greene, J. C. (2007). *Mixed methods in social inquiry.* San Francisco, CA: Jossey-Bass.

Greene, J.C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7-22.

Hesse-Biber, S. (2010). *Mixed methods research: Merging theory with practice*. New York: Guilford.

Hesse-Biber, S. (2012). Feminist approaches to triangulation: Uncovering subjugated knowledge and fostering social change in mixed methods research. *Journal of Mixed Methods Research, 6*(2), 137-146.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing*. Los Angeles: Sage.

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Labin, S. N., Duffy, J. L., Meyers, D. C., Wandersman, A., & Lesesne, C. A. (2012). A research synthesis of the evaluation capacity building literature. *American Journal of Evaluation, 33*(3), 307-338.

Mertens, D. M. (2005). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods*. Thousand Oaks, CA: Sage Publications.

Mertens, D. M., & Hesse-Biber, S. (2010). Triangulation and mixed methods research: Provocative positions. *Journal of Mixed Methods Research, 6*(2), 75-79.

Poole, D. L., Nelson, J., Carnahan, S., Chepenik, N. G., & Tubiak, C. (2000). Evaluating performance measurement systems in nonprofit agencies: The program accountability quality scale (PAQS). *American Journal of Evaluation, 21*(1), 15-26.

Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation, 29*(4), 443-459.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. doi: 10.1037/0033-2909.86.2.420

Stufflebeam, D. L. (1999). *Foundational models for 21st century program evaluation*: Evaluation Center, Western Michigan University.

Suarez-Balcazar, Y., Taylor-Ritzler, T., Garcia-Iriarte, E., Keys, C., Kinney, L., Rush-Ross, H., & Curtin, G. (2010). Evaluation capacity building: A cultural and contextual framework. *Race, culture and disability: Rehabilitation science and practice. Sudbury, MA: Jones & Bartlett Learning*.

Tessier, S. (2012). From field notes, to transcripts, to tape recordings: Evolution or combination? *International Journal of Qualitative Methods*, *11*(4), 446-460.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13*(1), 12-28.

Trochim, W., Urban, J. B., Hargraves, M., Hebbard, C., Buckley, J., Archibald, T., . . . Burgermaster, M. (2012). *The guide to the systems evaluation protocol*. Ithaca, NY: Cornell Digital Print Services.

Volkov, B., & King, J. A. (2007). A checklist for building organizational evaluation capacity. *The Evaluation Center: Western Michigan University. Accessed March, 28*, 2008.

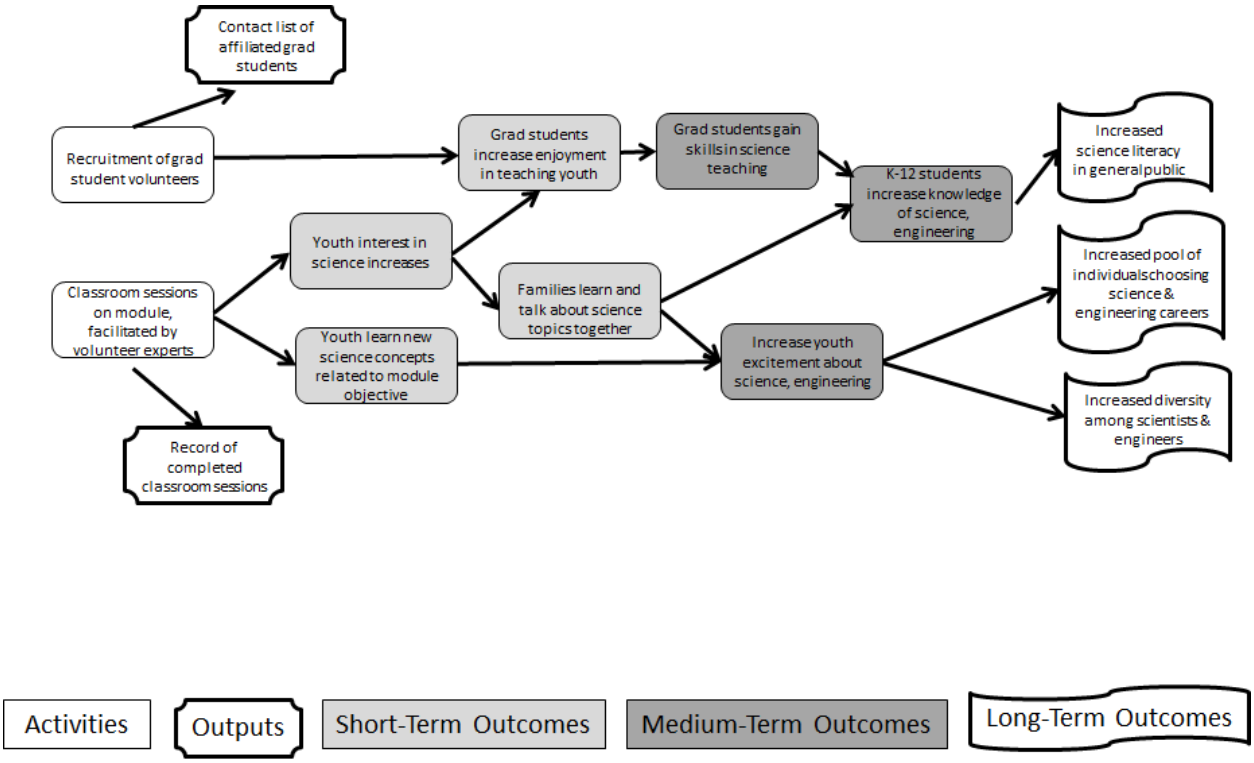

Figure 1. An example of a pathway model

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

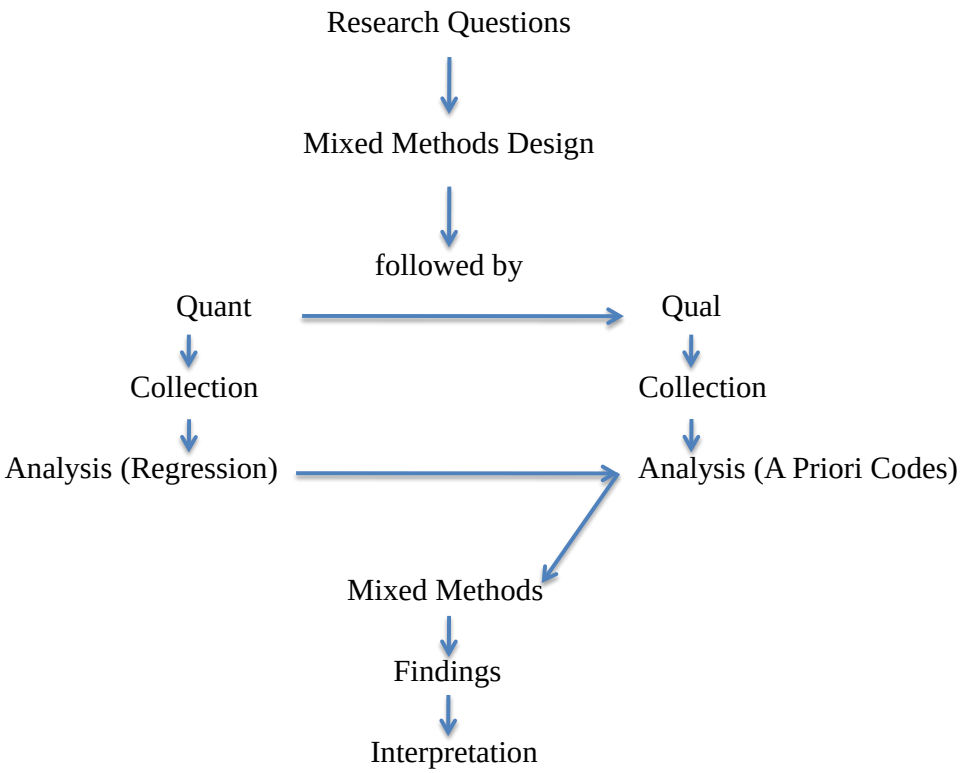

Figure 2. Modified quantitative → qualitative sequential explanatory design. In this modified quantitative → qualitative sequential explanatory design, the quantitative and qualitative data were collected based on the study's research questions. The research questions were used to develop a priori codes for the analysis of the qualitative data and a regression analysis was employed to analyze the quantitative data. The result of the regression analysis was then used to frame the analysis of the qualitative data. Finally, the mixed method findings were interpreted.

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

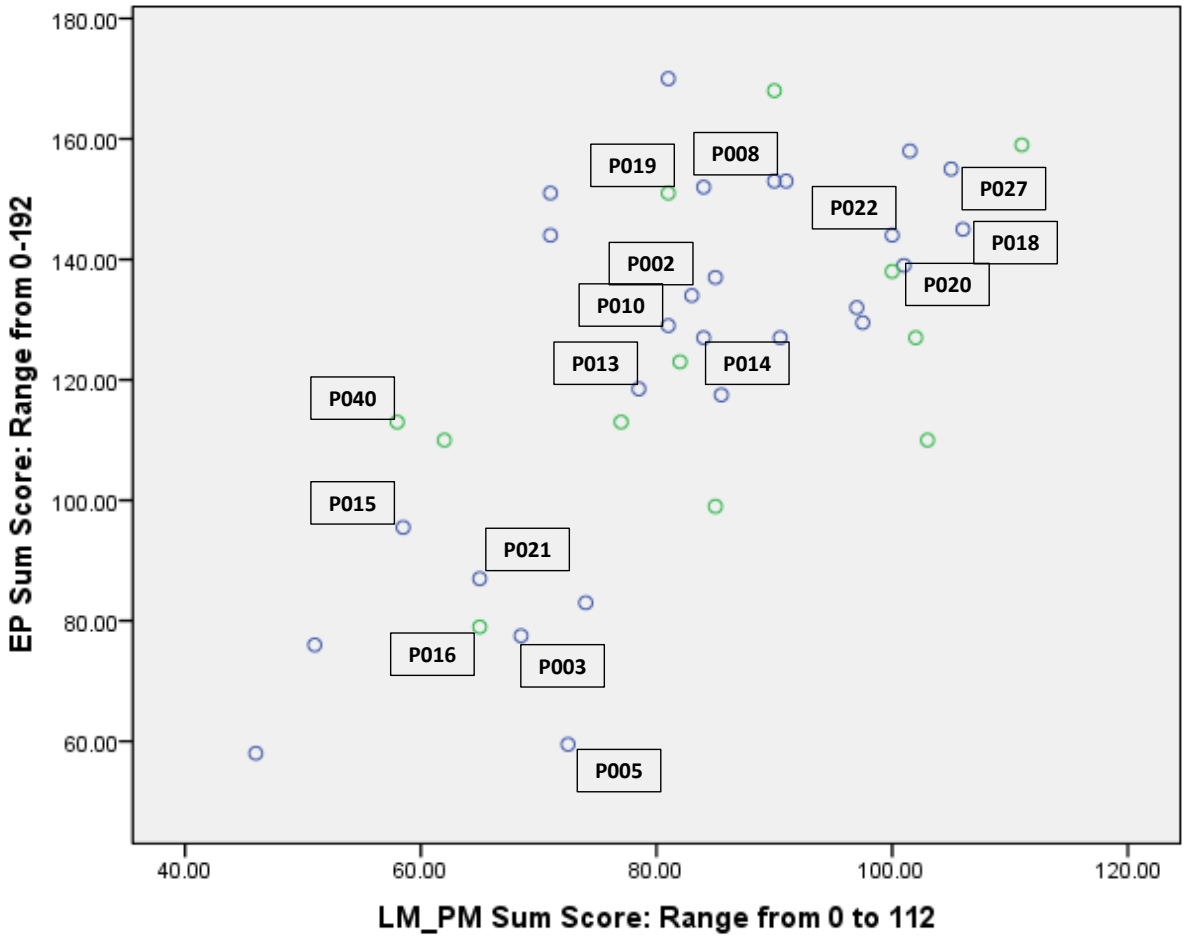

Figure 3. Logic Model and Pathway Model Sum Scores Predict Evaluation Plan Quality. Program codes (e.g., P016) indicate the program that is associated with a point on the regression line.

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Appendix A

Evaluation Plan Rubric

This rubric is intended to be a tool for reviewers to use in providing systematic scores on evaluation plans. In order to complete this rubric, the reviewer will need to have the evaluation plan, logic model, and/or pathway model for the target program.

Directions

1. Read through entire evaluation plan being reviewed.
2. Complete the scoring by section.
3. Score every item.
4. Items absent from the evaluation plan should be scored as "Unacceptable."
5. Items that are included in an evaluation plan but are not in the appropriate section (e.g., the measures are discussed in a section titled Design, and not in a section titled Measures) should still receive credit (and be scored in the appropriate section of the rubric).

General Guidelines

A good evaluation plan should

- provide an accurate, concise and coherent description of the program;
- explain what evaluation work is being planned and how the work will be accomplished;
- be appropriate for the program's content and stage of development; and
- be internally consistent (the elements of the evaluation plan should be consistent with each other - evaluation purpose, questions, measures, sampling strategy, design and analysis plans).

Section by Section Assessment

The categories below correspond to Evaluation Plan sections. Each item within a category provides a specific criterion for quality of work. The five-code scale is intended for numerical scoring. For each item, please refer to the Evaluation Plan (and logic model and/or pathway model as necessary) and check the box corresponding to your assessment of the item.

	Unacceptable	Minimally Acceptable	Adequate	Good	Excellent
Category/Criteria	**0**	**1**	**2**	**3**	**4**

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Program Mission or Purpose Statement					
1. Communication of goals (i.e., statement conveys the major goals of the program)					
2. Specificity to the program being evaluated (i.e., statement is about the program and *not* just the larger organization)					
Program Description					
3. Description of program implementation (e.g., includes information about target audience, program scale, activities, etc.)					
4. Description of program context (e.g., includes information about the social, cultural, physical context in which the program takes place)					
5. Description of intended outcomes or goals					
6. Description of program background (i.e., the history of the program's development is described and/or references to relevant research evidence base are included)					
Evaluation Purpose					
7. Identification of specific program activities, outcomes, or assumptions that are the focus of evaluation					
8. Articulation of the main goal(s) of the evaluation					
9. Description of intended use(s) of the evaluation					
10. Explanation of how the current evaluation plan fits in with any other (prior or ongoing) evaluation work on this program					
11. Appropriateness of evaluation goals relative to the stage of development of the program and its prior evaluation(s)					
Evaluation Questions					
12. Correspondence between questions and evaluation purpose (i.e., questions make sense given the evaluation purpose)					

13. Alignment between questions and the program's logic (i.e., questions are clearly related to the program's logic model and/or pathway model)					
14. Appropriateness of questions given the stage of development of the program and its prior evaluation(s)					
15. Feasibility of the set of questions (i.e., number of questions appears manageable)					
Sampling					
16. Alignment between sampling strategy and evaluation question(s) (i.e., each evaluation question is addressed)					
17. Description of population(s) of interest					
18. Description of sample(s)					
19. Description of recruitment for sample(s)					
20. Choice of sampling technique(s) (i.e., technique(s) such as simple random, convenience, cluster random, etc. will generate appropriate evidence)					
21. Appropriateness of sample size (i.e., sample size is sufficient for generating reasonable evidence; sample size is feasible)					
Measurement					
22. Alignment between measurement strategy and evaluation question(s) (i.e., each evaluation question is addressed)					
23. Description of measures (i.e., description of each measure is clear and includes type of measure — e.g., survey, observation, interview, etc.)					
24. Appropriateness of selected type of measure (e.g., survey, observation, interview, etc.) for generating evidence to answer evaluation question(s)					
25. Appropriateness of measures for program setting, audience, etc.					

26. Identification of focal construct/variable for each measure (i.e., the construct or variable that will be measured is clearly identified)					
27. Description of origin or development of each measure (e.g., appropriately cited, development of new tool described)					
28. Description of measure quality (i.e., validity and reliability issues are appropriately addressed given the stage of development of the program and its evaluation)					
Design					
29. Alignment between design strategy and evaluation questions (i.e., each evaluation question is addressed)					
30. Description of design (e.g., post-only, pre/post, pre/post with comparison group, etc.)					
31. Appropriateness of selected design(s) given the stage of development of the program and its prior evaluation(s)					
32. Appropriateness of selected design(s) for generating evidence to answer evaluation question(s)					
Data Collection and Management					
33. Alignment between data collection strategy/management and evaluation questions or measures (i.e., either each evaluation question is addressed or each measure is addressed)					
34. Comprehensiveness (i.e., addresses data collection and management steps including measure(s) administration, data capture, data handling, data storage, etc.)					
35. Plan for how data will be organized in preparation for analysis					
Data Analysis					
36. Alignment between data analysis strategy and evaluation question(s) (i.e., each evaluation question is addressed)					
37. Appropriateness of data analysis strategies for generating evidence to answer evaluation question(s)					

Evaluation Reporting and Utilization					
38. Alignment between evaluation reporting/utilization strategy and evaluation question(s) (i.e., each evaluation question is addressed)					
39. Comprehensiveness (e.g., addresses external and internal reporting, the form and frequency of reporting, the intended uses such as feedback to staff, program improvement, accountability)					
40. Appropriateness of plans for utilizing evaluation results given the evaluation purpose					
41. Appropriateness of plans for utilizing evaluation results relative to the program's current stage of development					
Evaluation Timeline					
42. Inclusion of program and/or activity events					
43. Inclusion of evaluation plan implementation events (e.g., measures development/procurement, data collection, etc.)					
44. Specificity (i.e., timeline events are given in calendar time, not just in relative terms; clear start and end dates are provided)					
Overall					
45. Quality of writing (i.e., clarity, consistency of voice, correct grammar, proper spelling)					
46. Quality of evaluation plan as communication tool (i.e., language and phrasing are understandable to outside readers; does not use program-specific terms, acronyms)					
47. Internal alignment (i.e., sample, design, measurement, analysis plans are consistent with evaluation questions *and* with each other)					
48. Feasibility of Evaluation Plan					

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Appendix B

Logic & Pathway Model Rubric

This rubric is intended to be a tool for external reviewers to use in providing systematic scores on logic models and pathway models. In order to complete this rubric, the reviewer will need to have the evaluation plan, logic model, and pathway model for the target program.

Directions

1. Review the program as described in the evaluation plan.
2. Review the program's logic model and complete the rubric for logic models.
3. Review the program's pathway model and complete the rubric for pathway models.
4. Items absent from the logic and/or pathway models should be scored as "Unacceptable."

Logic Model General Guidelines

A good logic model should

- describe a program accurately, concisely, and coherently;
- reflect the internal logic of the program;
- be consistent with the program as described in the evaluation plan; and
- present a plausible program logic that is internally consistent.

Logic Model Section by Section Assessment

The categories below correspond to Logic Model sections. Each item within a category provides a specific criterion for quality of work. The scale is intended for numerical scoring. For each item, please refer to the logic and pathway models and check the box corresponding to your assessment of that item.

	Unacceptable	Minimally Acceptable	Adequate	Good	Excellent
Category/Criteria	**0**	**1**	**2**	**3**	**4**
Inputs					
1. Phrasing of inputs (e.g., language and phrasing are understandable to outside readers; does not use program-specific terms, acronyms)					
2. Summary of program size/scale (e.g., provides one or more of: % FTE for staff; annual budget; average number of participants, etc. as appropriate)					

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Activities					
3. Phrasing of activities (e.g., language and phrasing are understandable to outside readers; does not use program-specific terms, acronyms)					
4. Consistency of activities with program as described in the evaluation plan					
5. Appropriateness of activities (e.g., list only includes activities that reach people who participate or who are targeted)					
Outputs					
6. Phrasing of outputs (e.g., language and phrasing are understandable to outside readers; does not use program-specific terms, acronyms)					
7. Comprehensiveness of outputs list (i.e., outputs are included for activities that are likely to generate outputs)					
8. Appropriateness of outputs (i.e., they are closely linked by-products of program activities and do *not* include effects on participants; depending on program, outputs might include: attendance lists, certificates of completion, projects completed, etc.)					
Outcomes					
9. Phrasing of outcomes (e.g., language and phrasing are understandable to outside readers; does not use program-specific terms, acronyms)					
10. Consistency of outcomes with program as described in the evaluation plan (i.e., outcomes reflect the scope of the program)					
11. Appropriateness of outcomes (i.e., outcomes are phrased as effects on, or changes in, participants and/or their communities or society and are *not* actions, objectives, or specific indicators)					
12. Placement of outcomes (i.e., outcomes are in the correct columns; they logically arise with or "soon after" the preceding activities or outcomes)					
Assumptions					
13. Phrasing of assumptions (e.g., language and phrasing are understandable to outside readers; does not use program-specific terms, acronyms)					
14. Description of assumptions (e.g., assumptions describe beliefs and thinking about the program and how it will occur)					

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388

Context					
15. Phrasing of program context (e.g., language and phrasing are understandable to outside readers; does not use program-specific terms, acronyms)					
16. Description of context (e.g., describes the social, cultural, physical context in which the program is taking place)					
Overall					
17. Quality of writing (i.e., clarity, correct grammar, proper spelling)					
18. Quality of model as communication tool (i.e., provides a reasonable and understandable summary of the program)					

Pathway Model General Guidelines

- A good pathway model should be comprehensive and internally consistent. However, there is no prescriptive level of detail or generality that is "right" under all circumstances.

  Special note regarding the "Connections" section: Pathway models are inherently inter-connected and therefore, you may find that there is overlap in the items listed in the "Connections" section of the rubric. For example, a model that has a short-term outcome connected directly to a long-term outcome should be given a lower score on ***both* item 2 (sequencing of connections to and from short-term outcomes) and item 4 (sequencing of connections to and from long-term outcomes).

Pathway Model Section by Section Assessment

The categories below correspond to Pathway Model elements. Each item within a category provides a specific criterion for quality of work. The five-code scale is intended for numerical scoring. For each item, please refer to the pathway model and check the box corresponding to your assessment of that item.

	Unacceptable	**Minimally Acceptable**	**Adequate**	**Good**	**Excellent**

Category/Criteria	0	1	2	3	4
Items					
1. Consistency between pathway and logic model (i.e., activities and outcomes from logic model are present in pathway model)					
Connections					
2. Sequencing of connections to and from short-term outcomes (i.e., they are appropriately connected to activities and other medium-term or short-term outcomes, but *not* directly to long-term outcomes)					
3. Sequencing of connections to and from medium-term outcomes (i.e., they are appropriately connected to other outcomes but *not* to activities)					
4. Sequencing of connections to and from long-term outcomes (i.e., they are appropriately connected to medium-term or other long-term outcomes but *not* directly to short-term outcomes or activities)					
5. Completeness of connections (i.e., connections that should be made are made)					
6. Plausibility of connections (i.e., the connections make sense and are logical)					
Pathways (refers to explanatory "through lines" that connect specific activities and outcomes)					
7. Logic of pathways (i.e., they communicate the "story" or "theory" that joins activities to long term outcomes)					
8. Completeness of pathways (i.e., they do not dead-end at outputs, short- or mid-term outcomes, but rather follow through to long-term outcomes)					
Overall					
9. Quality of pathway diagram as communication tool (i.e., diagram efficiently communicates overall program logic)					
10. Readability of model (e.g., number of connections is not excessive)					

Urban, J.B., Burgermaster, M., Archibald, T., & Byrne, A. (2015). Relationships between quantitative measures of evaluation plan and program model quality and a qualitative measure of participant perceptions of an evaluation capacity building approach. *Journal of Mixed Methods Research, 9*(2), 154-177. DOI: 10.1177/1558689813516388